Predição de Alunos em Risco de Reprovação: uma Comparação do Impacto de Diferentes Técnicas de Amostragem

Júlio César da Silva Dantas, UFRN, dantas.j.c.s@gmail.com, https://orcid.org/0000-0002-3729-9662.
Eduardo Henrique da Silva Aranha, UFRN, eduardoaranha@dimap.ufrn.br, https://orcid.org/0000-0002-8446-638X.
Thiago Reis da Silva, IFMA/Campus São João dos Patos, thiago.reis@ifma.edu.br, https://orcid.org/0000-0003-4206-6801.

Resumo: *Entre os campos impactados pela Inteligência Artificial, a educação é um dos mais transformados, com aplicações na automatização de processos, suporte em atividades e, como neste caso, na Predição de Performance de Estudantes (student performance prediction, SPP). Esta pesquisa investiga a eficácia de técnicas de amostragem, como SMOTE, Tomek Links e ADASYN, no problema de desequilíbrio de classes, comum na área de SPP. Trata-se de uma pesquisa aplicada, exploratória e experimental, que alcançou 97,5% de precisão com o uso de Gradient Boosting e SMOTE, identificando até 94,4% dos alunos em risco de reprovação. Esses resultados superam trabalhos anteriores com o mesmo conjunto de dados, sugerindo a eficiência das técnicas de amostragem para lidar com o problema de desequilíbrio de classes.*
Palavras-chave: *Predição de performance de estudantes, SMOTE, Tomek links, ADASYN.*

Prediction of Students at Risk of Failure: A Comparison of the Impact of Different Sampling Techniques

ABSTRACT: *Among the fields impacted by Artificial Intelligence, education is one of the most transformed, with applications in process automation, activity support, and, as in this case, Student Performance Prediction (SPP). This research investigates the effectiveness of sampling techniques such as SMOTE, Tomek Links, and ADASYN in addressing class imbalance, a common issue in the SPP domain. This is an applied, exploratory, and experimental study, achieving 97.5% accuracy using Gradient Boosting and SMOTE, identifying up to 94.4% of students at risk of failure. These results surpass previous studies with the same dataset, suggesting the efficiency of sampling techniques in handling class imbalance issues.*
Keywords: *Student performance prediction, SMOTE, Tomek links, ADASYN.*

1. Introdução

Identificar alunos em risco de reprovação, em especial no contexto presencial da educação básica, é um desafio para professores e pessoal das instituições de ensino: seja porque têm muitos estudantes sob sua tutela, seja porque os sinais de dificuldade nem sempre são evidentes o suficiente. Antecipar a reprovação de um estudante significa ter mais oportunidades de oferecer suporte e, no melhor dos casos, evitar a concretização da previsão de reprovação.

Neste contexto, diversas soluções de aprendizado de máquina, do inglês *Machine Learning* (ML), são utilizadas na literatura para a previsão de alunos em risco de reprovação. Essas soluções podem se basear na performance do estudante em Ambientes Virtuais de Aprendizagem (AVAs), como em Stoll, Cury e Menezes (2018); no histórico de desempenho, como em Filho *et al*. (2021); ou, em sua forma mais comum, em dados demográficos dos estudantes, como apresentando em Neo *et al.* (2023).

Modelos erguidos em dados de AVAs são os que, geralmente, apresentam melhor desempenho; para isso, os modelos coletam dados de cliques, número de vídeos assistidos, atividades realizadas, dentre outras características. Com o extenso conjunto de dados e a complexidade dos modelos empregados, essas soluções apresentam baixa interpretabilidade e comprometem a replicabilidade, visto que são limitadas a contextos que utilizam AVAs – uma realidade pouco comum na maior parte da educação básica brasileira.

Modelos que usam o histórico de notas dos estudantes nos anos anteriores apresentam boas métricas de desempenho, mas isso exige uma rigorosa estrutura de organização de dados, talvez precisando contar também com a cooperação de outras instituições de ensino e, assim como modelos baseados em dados demográficos discutidos adiante, não conseguem capturar a evolução dos estudantes com o passar do tempo.

Outra fonte de dados usada na literatura é da demografia do público, incluindo informações sobre, por exemplo, a profissão e educação dos pais, renda e gênero. Esses modelos também carecem de cuidado especial quanto à privacidade dos dados, não capturam mudanças correntes nos perfis dos estudantes e correm o risco de incorrer em preconceitos, a depender do conjunto de dados usado.

Independente da origem do conjunto de dados, um obstáculo comum no problema de identificar estudantes em risco de reprovação por meio de técnicas de aprendizado de máquina é o desequilíbrio entre as classes (Imran *et al*. 2023). Em outras palavras, como há significativamente mais estudantes aprovados do que reprovados, os modelos têm dificuldade em aprender as características da classe minoritária, resultando em um baixo desempenho na identificação dessa classe durante os testes, conforme observado também por Bujang *et al.* (2019).

Sendo assim, esta pesquisa tem como um dos objetivos investigar o desempenho de diferentes técnicas de amostragem nas métricas de avaliação de modelos de inteligência artificial (IA) empregados no problema de predição de performance de estudantes, em atenção à lacuna identificada na literatura, que sugere a deterioração da avaliação desses modelos na presença de desequilíbrio entre as classes. Para isso, deve coletar e processar os dados disponibilizados por Cortez e Silva (2008) e desenvolver e avaliar os modelos de ML, com e sem o uso de técnicas de amostragem. Deve responder às questões: *técnicas de amostragem têm impacto significativo no desempenho de modelos de inteligência artificial no contexto de predição de performance de estudantes?* Se sim, qual combinação modelo-amostragem tem melhor desempenho no conjunto disponível?

As próximas sessões do trabalho discutem os pesquisas relacionadas, em especial aqueles que usam o mesmo conjunto de dados; a definição teórica das técnicas de amostragem propostas, o contorno metodológico desta pesquisa, os resultados dos experimentos computacionais e as considerações finais, com as principais contribuições desta pesquisa e a propostas de trabalhos futuros.

2. Material e Métodos

Por gerar conhecimentos a partir da aplicação de teorias, esse trabalho pode ser classificado como aplicado; para isso, determinamos um objeto de estudo, selecionamos as variáveis que seriam capazes de influenciá-lo, definimos as formas de controle e de observação dos efeitos que a variável produz no objeto, fazendo que se enquadre sob o paradigma experimental (Prodanov e Freitas, 2013). Porque constrói suas conclusões da análise numérica do desempenho dos modelos, é também quantitativa. Além disso, por inaugurar a investigação dos impactos das técnicas de amostragem no contexto da

predição de desempenho estudantil e gerar hipóteses para estudos futuros, ela também pode ser classificada como exploratória (Prodanov e Freitas, 2013).

Os experimentos foram realizados usando as bibliotecas *numpy* (2.1.1), *imblearn* (0.12.4), *pandas* (2.2.2), *scikit-learn* (1.5.1) e *xgboost* (2.1.1). Todos os modelos de ML foram implementados usando as versões da biblioteca *scikit-learn*, exceto *XGBoost*. Os dados foram divididos entre treinamento e teste garantindo a proporção entre classes da variável alvo, validados com 10 *folds*. Variáveis contínuas sofreram *scaling* (*MinMaxScaler*), categóricas sofreram *encoding*, mas nenhuma *feature engineering* e *feature selection* é feita e, sendo assim, os modelos contam com o mesmo número de variáveis que o *dataset* original.

O conjunto de dados disponibilizado por Cortez e Silva (2008), junto com OULAD, são dois amplamente utilizados na literatura no problema de predição de performance dos estudantes; no quadro abaixo, alguns desses trabalhos.

Tabela 1: Estudos que usam o dataset coletado por Cortez e Silva (2008).

Trabalho	Métricas alcançadas
Ali, Aborizk, Dharoug (2023)	CNN + RNN: *Accuracy*: 0.891, *precision*: 0.887, *recall*: 0.907, *f1-score*: 0.897.
Muruganandam, Rajini (2021)	DNN + AAFO: *Accuracy*: 0.979, *recall* 0.954.
Chowdhury *et al.* (2022)	Random Forest: *Accuracy*: 0.933.
Ram *et al.* (2021)	Random Forest: *Accuracy*: 0.82, *precision*: 0.77, *recall*: 0.87, *f1-score*: 0.82.
Razak *et al.* (2021)	Decision Tree: *Accuracy*: 0.506, *precision*: 0.628, *recall*: 0.577, *f1-score*: 0.606.
Assistant *et al.* (2021)	Decision table: *Accuracy*: 0.761, *precision*: 0.771, *recall*: 0.761.
Roy, Garg (2017)	Naive-Bayes: *Precision*: 0.894, *recall*: 0.715.
Kiu (2018)	Random Forest: *Accuracy*: 0.924.

Fonte: Elaborados pelos autores (2024).

O conjunto de dados apresenta 32 features, com 649 observações, divididas entre os resultados da disciplina de matemática e português - sendo apenas o primeiro utilizado no estudo. Entre as variáveis constam o sexo do estudante, idade, endereço, situação de casamento dos pais, educação e trabalho dos pais, parente responsável pela criança, tempo de estudo livre e de viagem a escola, atividades extracurriculares, consumo de álchool, notas do primeiro e segundo semestres e nota final - variável a que se deseja antecipar. Os dados foram coletados entre 2006 e 2007, de duas escolas públicas de Alentejo, Portugal. Os instrumentos de coleta foram relatórios da escola e questionários (Cortez e Silva, 2008).

3. Resultados e Discussão

3.1 Estado das Pesquisas no Brasil

No Brasil, alguns trabalhos tiveram proposta semelhante: Stoll, Cury e Menezes (2018) usam dados de ambientes virtuais de aprendizagem (AVAs) na predição e propõem um sistema com modelos de ML em que a escolha do modelo se dá a partir do desempenho no conjunto de dados submetido ao sistema. Os autores atingem a melhor performance com *Decision tree*, chegando a 0.702 de acurácia.

Santana, Gusmão e Gusmão (2023) tiveram o objetivo de prever o resultados dos estudantes em uma avaliação de português a partir de informações associadas às condições socioeconômicas dos alunos e situação estrutural das escolas. Os autores usaram dados do Sistema Nacional de Avaliação da Educação Básica (SAEB), realizaram

feature selection e *oversampling* da classe minoritária, alcançando, com *Random forest*, acurácia de 81%, precisão de 83%, *recall* de 97% e *f-score* de 89%.

Filho *et al.* (2021) focam no Exame Nacional do Ensino Médio (ENEM) e identificaram correlação relevante das notas durante o terceiro ano do ensino médio e a nota no ENEM; tratam como um problema de regressão, e não classificação, e usam *SimpleLinearRegression* do Weka para atingir *root mean squared error* de 40 pontos.

Neo *et al*. (2023) se baseiam em dados demográficos no contexto do ensino técnico e atingem *f-score* de 64% com *Naive-Bayes*, com apenas 25 instâncias. Gottardo, Kaestner e Noronha tematizam a modalidade de Educação a Distância, usam dados minerados de AVA com 140 participantes, atingindo acurácia entre 80-73% com *Random Forest e MLPClassifier*.

Costa *et al.* (2017) em contrapartida, analisa dados da educação a distância e de interações presenciais, além dos demográficos. Os autores têm preocupação de investigar a predição em diferentes momentos do curso, usando diferentes conjuntos de dados. Como conclusão, chegam ao *support vector machine* como modelo que melhor performa, atingindo *f-score* de 0.92.

Este estudo também se baseia em dados demográficos dos estudantes na antecipação de reprovação, mas se distancia dos outros trabalhos na proposta do uso e comparação de técnicas de amostragem para endereçar o problema de desequilíbrio de classes.

3.2. Técnicas De Amostragem Utilizadas

Uma das abordagens para lidar com conjuntos de dados desbalanceados é modificar a distribuição das classes, tornando-a mais equilibrada. Existem, pelo menos, dois métodos para isso: *undersampling* e *oversampling*. O primeiro consiste em reduzir o número de elementos da classe majoritária, enquanto o segundo envolve aumentar o número de elementos da classe minoritária (Batista, Bazzan e Monard, 2003).

Neste trabalho, foram utilizadas quatro propostas de amostragem: *Synthetic Minority Over-sampling Technique* (SMOTE), como definida em Chawla *et al.* (2002), Tomek links, uma técnica de *undersampling*, uma técnica mista de SMOTE + Tomek links, como definida em Batista, Bazzan e Monard (2003) e *Adaptive Synthetic Sampling Approach for Imbalanced Learning* (ADASYN), como definida em He *et al.* (2008). Os métodos são definidos adiante.

Synthetic Minority Over-sampling Technique usa *oversampling* mas, para além de copiar dados já existentes na classe minoritária, sintetiza novos pontos a partir de k vizinhos mais próximos. Mais especificamente, calcula-se a diferença entre o vetor de características em consideração e seu vizinho mais próximo, multiplica-se essa diferença por um número aleatório entre 0 e 1, e adiciona o resultado ao vetor de características em consideração. Isso resulta na seleção de um ponto aleatório ao longo do segmento de linha entre duas características específicas. Esse método, na prática, força a região de decisão da classe minoritária a se tornar mais generalizada (Chawla *et al*., 2008). Assim, SMOTE pode melhorar consideravelmente o desempenho dos modelos, mas é especialmente vulnerável a *outliers*.

Undersampling e *oversampling* possuem desvantagens: o primeiro pode desperdiçar informação e o segundo pode aumentar a possibilidade de *overfitting* do modelo. Sendo assim, Batista, Bazzan e Monard (2003), três pesquisadores brasileiros, propõem aliar a SMOTE à técnica de *undersampling* Tomek links: enquanto que o primeiro cria mais instâncias da classe minoritária, o que pode ajudar ao modelo a identificá-lá, o segundo remove algumas instâncias (de ambas as classes), de maneira a deixar os *clusters* de classes mais claros. Um *Tomek link* pode ser definido assim: dados

dois exemplos, x e y, pertencentes a classes diferentes, e sendo $d(x, y)$ a distância entre x e y, um par (x, y) é chamado de Tomek link se não houver um caso z, tal que $d(x, z) < d(x, y)$ ou $d(y, z) < d(y, x)$ (Batista, Bazzan, Monard, 2003).

O método *Adaptive Synthetic Sampling Approach for Imbalanced Learning* (ADASYN) também se baseia na síntese de novos pontos para a classe minoritária, utilizando uma abordagem que gera amostras adaptativamente, conforme a distribuição dos dados. Mais dados sintéticos são criados para os exemplos da classe minoritária que são mais difíceis de aprender, enquanto menos amostras são geradas para aqueles que são mais fáceis de aprender. O ADASYN não apenas reduz o viés de aprendizado causado pelo desequilíbrio original na distribuição dos dados, mas também ajusta adaptativamente o limite de decisão, focando nas amostras que apresentam maior dificuldade de aprendizado.

A diferença entre ADASYN e SMOTE está no critério de geração das amostras sintéticas: enquanto o SMOTE cria novas amostras de forma uniforme, gerando pontos ao longo da linha entre uma amostra minoritária e seus vizinhos mais próximos, sem considerar a complexidade das amostras, o ADASYN adapta a geração de dados com base na dificuldade de aprendizado, e foca em gerar mais amostras sintéticas para os exemplos minoritários que são mais difíceis de classificar, ajustando o limite de decisão do modelo para essas amostras complexas, enquanto o SMOTE busca apenas balancear a distribuição das classes de maneira geral.

3.3. Resultados dos Experimentos

A seguir constam os resultados obtidos pelos modelos, divididos em duas partes. Na primeira parte, constam duas *features* correlacionadas com a variável que se pretende prever. São elas: G1 e G2, que são a nota dos estudantes na primeira e segunda partes do semestre. Na segunda parte destes resultados, essas notas não integram o conjunto e a razão disso é discutida adiante. Para cada parte, há duas tabelas: em uma constam os valores de *accuracy* (acc), *precision* (pre) e *recall* (rec) para cada combinação de modelo-técnica de amostragem. As colunas de *precision* e *recall* trazem a média ponderada dessas grandezas em relação ao número de instâncias de cada classe no conjunto de testes, como é apropriado em um problema de *class imbalance*. Na segunda tabela constam os valores de *recall* alcançados em cada classe (apr, aprovado e rep, reprovado), de maneira a mitigar outro problema com desequilíbrio de classes: um modelo pode atingir altos valores de *accuracy* apenas prevendo a classe majoritária para todas ou a maioria dos casos de teste. Nesse sentido, é importante conhecer as métricas de avaliação dentro das classes, em especial neste contexto de aplicação, em que um falso positivo para a classe de reprovação é menos importante que um falso negativo.

A escolha por realizar dois experimentos, um contendo as notas anteriores dos estudantes e outro sem essas variáveis, se justifica por dois fatores. Primeiro, como o objetivo deste trabalho é comparar o desempenho das técnicas de amostragem com outros estudos na literatura que utilizam essas variáveis, é necessário partir de uma base comum, o que justifica o uso das notas. No entanto, ao utilizar o resultado de testes realizados ao longo da disciplina, perde-se o foco nas variáveis demográficas, que têm o potencial de identificar alunos em risco de reprovação antes do início da disciplina. Por essa razão, uma versão dos modelos sem essas variáveis no conjunto de dados também é investigada.

Dos resultados apresentados na Tabela 2, é possível perceber que as técnicas de amostragem melhoram o desempenho dos modelos em até três pontos percentuais entre as métricas de avaliação, o que é relevante para modelos que já alcançam alta taxa de precisão. As técnicas de amostragem são especialmente relevantes para os modelos

baseados em *boosting*, como *GradientBoosting*, *XGBoost* e *AdaBoost*, com o primeiro sendo o que melhor performa, combinado com SMOTE, nas métricas gerais.

Tabela 2: Resultados da avaliação dos modelos, com as features *G1 e G2* no conjunto de dados.

Modelos	**Base**			**SMOTE**			**SMOTE + Tk**			**Tomek**			**ADASYN**		
	acc	**pre**	**rec**	**acc**	**pre**	**rec**	**acc**	**pre**	**rec**	**acc**	**pre**	**rec**	**acc**	**pre**	**rec**
SVC	0.866	0.91	0.866	0.908	0.906	0.908	0.908	0.906	0.908	0.84	0.902	0.84	0.899	0.899	0.899
RandomForest	**0.941**	0.939	**0.941**	0.941	0.94	0.941	0.95	0.948	0.95	0.958	0.96	0.958	**0.958**	0.957	**0.958**
GradientBoosting	**0.941**	**0.943**	**0.941**	**0.975**	**0.976**	**0.975**	**0.975**	**0.976**	**0.975**	**0.95**	0.953	0.95	**0.958**	**0.959**	**0.958**
MLPClassifier	0.908	0.906	0.908	0.899	0.895	0.899	0.899	0.895	0.899	0.908	0.906	0.908	0.908	0.903	0.908
LogisticRegression	0.916	0.932	0.916	0.916	0.92	0.916	0.916	0.92	0.916	0.899	0.924	0.899	0.916	0.92	0.916
XGBoost	**0.941**	**0.943**	**0.941**	0.941	0.943	0.941	0.941	0.943	0.941	0.941	0.943	0.941	**0.958**	**0.959**	**0.958**
Naive-bayes	0.84	0.893	0.84	0.832	0.872	0.832	0.832	0.872	0.832	0.84	0.893	0.84	0.832	0.864	0.832
KNNeighbors	0.824	0.717	0.824	0.689	0.772	0.689	0.689	0.772	0.689	0.824	0.717	0.824	0.739	0.786	0.739
AdaBoost	**0.941**	**0.943**	**0.941**	0.941	0.943	0.941	0.941	0.943	0.941	0.966	**0.966**	**0.966**	0.95	0.95	0.95

Fonte: Elaborada pelos autores.

Os dados da Tabela 3 corroboram os resultados pontuados sobre a efetividade das técnicas de amostragem e deixam evidente a capacidade de ajudar a identificar a classe minoritária. Enquanto os modelos *baseline* conseguiram identificar até 0.889 da classe de reprovação no conjunto de teste, *GradientBoosting*, associado com SMOTE, conseguiu identificar até 0.944 - um aumento de 0.055. Comparando o mesmo modelo antes e depois das técnicas de amostragem, o aumento da precisão na identificação da classe minoritária chega a mais de 10 pontos percentuais.

Tabela 3: Resultados de *recall* para as classes, com as features G1 e G2 no conjunto de dados.

Modelos	**Base**		**SMOTE**		**SMOTE + Tk**		**Tomek**		**ADASYN**	
	apr	**rep**	**apr**	**rep**	**apr**	**rep**	**apr**	**rep**	**apr**	**rep**
SVC	0.861	**0.889**	0.95	0.667	0.95	0.667	0.832	**0.889**	0.941	0.667
RandomForest	**0.98**	0.722	0.97	0.778	**0.98**	0.778	1	0.722	**0.98**	0.833
GradientBoosting	0.96	0.833	**0.98**	**0.944**	**0.98**	**0.944**	0.96	**0.889**	0.97	**0.889**
MLPClassifier	0.95	0.667	0.95	0.611	0.95	0.611	0.95	0.667	0.96	0.611
LogisticRegression	0.921	**0.889**	0.941	0.778	0.941	0.778	0.901	**0.889**	0.941	0.778
XGBoost	0.96	0.833	0.96	0.833	0.96	0.833	0.96	0.833	0.97	**0.889**
Naive-bayes	0.842	0.833	0.851	0.722	0.851	0.722	0.842	0.833	0.861	0.667
KNNeighbors	0.97	0	0.743	0.389	0.743	0.389	0.97	0	0.802	0.389
AdaBoost	0.96	0.833	0.96	0.833	0.96	0.833	**0.98**	**0.889**	0.97	0.833

Fonte: Elaborada pelos autores.

No contexto do conjunto de dados menos informativo, na ausência das variáveis *G1* e *G2*, os resultados da Tabela 4 fazem parecer, à primeira vista, que as técnicas de amostragem não foram relevantes no desempenho dos modelos. Entre os resultados, a combinação de *Naive-Bayes* e Tomek link fez passar de 0.891 de *accuracy* para 0.908 Na verdade, alguns modelos como o *MLPClassifier*, *LogisticRegression* e *KNN* apresentam piora na performance; apenas SVC tem melhora expressiva, embora ainda com menor desempenho que outros modelos.

Tabela 4: Resultados da avaliação dos modelos, com o conjunto de dados sem as *features G1 e G2*.

Modelos	**Base**			**SMOTE**			**SMOTE + Tk**			**Tomek**			**ADASYN**		
	acc	**pre**	**rec**	**acc**	**pre**	**rec**	**acc**	**pre**	**rec**	**acc**	**pre**	**rec**	**acc**	**pre**	**rec**
SVC	0.151	0.023	0.151	0.807	0.772	0.807	0.807	0.772	0.807	0.151	0.023	0.151	0.748	0.769	0.748
RandomForest	0.849	0.72	0.849	0.849	0.806	0.849	0.849	0.806	0.849	0.849	0.72	0.849	**0.874**	**0.858**	**0.874**
GradientBoosting	0.857	0.833	0.857	**0.866**	**0.844**	**0.866**	**0.866**	**0.844**	**0.866**	0.866	0.851	0.866	0.857	0.829	0.857
MLPClassifier	0.849	0.72	0.849	0.79	0.775	0.79	0.773	0.744	0.773	0.79	0.775	0.79	0.782	0.76	0.782
LogisticRegression	0.849	0.72	0.849	0.807	0.793	0.807	0.807	0.793	0.807	0.798	0.834	0.798	0.807	0.793	0.807
XGBoost	0.849	0.72	0.849	0.84	0.808	0.84	0.84	0.808	0.84	0.849	0.72	0.849	0.815	0.777	0.815
Naive-bayes	**0.891**	**0.887**	**0.891**	0.723	0.762	0.723	0.723	0.762	0.723	**0.908**	**0.906**	**0.908**	0.689	0.752	0.689
KNNeighbors	0.84	0.719	0.84	0.664	0.746	0.664	0.664	0.746	0.664	0.824	0.783	0.824	0.655	0.753	0.655
AdaBoost	0.849	0.72	0.849	0.84	0.815	0.84	0.84	0.815	0.84	0.849	0.72	0.849	0.815	0.798	0.815

Fonte: Elaborada pelos autores.

A análise dos dados em termos do *recall* das classes aprovado/reprovado, no entanto, deixa claro que as técnicas tiveram impacto relevante na predição, em especial da classe minoritária. A piora no desempenho geral de alguns modelos está associada não ao prejuízo das técnicas de amostragem, mas ao fato que os modelos passaram a não mais prever apenas a classe majoritária para todas as instâncias de teste, como discutido melhor adiante.

Usando o conjunto de dados menos informativo, todos os modelos com o conjunto de dados *baseline*, exceto *Naive-Bayes*, sofreram com um problema comum no contexto de *class imbalance* conhecido na literatura como *class collapse*, em que o modelo prediz uma única classe para todos os casos de teste. Esse fenômeno também aconteceu com alguns modelos em comunhão com a técnica de *undersampling* Tomek links e pode ser explicado como se segue: com um conjunto de dados menos informativo, a retirada de observações deixou o conjunto ainda menos informativo, levando ao *class collapse*.

Sendo assim, dos dados é possível perceber que os modelos só conseguem apreender as características do conjunto de treinamento a partir das técnicas de amostragem, em especial a SMOTE para modelos mais complexos e *Tomek links* para modelos mais simples. Entre os modelos que melhor performam na identificação da classe minoritária no conjunto de dados menos informativo, cabe citar *Naive-Bayes*, identificando 0.990 da classe majoritária e 0.444 da minoritária, e *LogisticRegression*, que identifica 0.842 dos aprovados e 0.556 dos reprovados; entre eles, *Naive-Bayes*,

aliado a Tomek links, é o que tem a performance mais equilibrada, que reflete nas melhores métricas gerais de desempenho da tabela 4.

Tabela 5: Resultados de recall para as classes, com o conjunto de dados sem as features *G1* e *G2*.

Modelos	Base		SMOTE		SMOTE + Tk		Tomek		ADASYN	
	apr	**rep**	**apr**	**rep**	**apr**	**rep**	**apr**	**rep**	**apr**	**rep**
SVC	0	1	0.921	0.167	0.921	0.167	0	1	0.832	0.278
RandomForest	1	0	0.98	0.111	0.98	0.111	1	0	0.98	0.278
GradientBoosting	0.99	0.111	0.98	0.222	0.98	0.222	0.99	0.167	0.97	0.222
MLPClassifier	1	0	0.891	0.222	0.891	0.111	0.891	0.222	0.891	0.167
LogisticRegression	1	0	0.901	0.278	0.901	0.278	0.842	**0.556**	0.901	0.278
XGBoost	1	0	0.95	0.222	0.95	0.222	1	0	0.931	0.167
Naive-bayes	0.99	0.333	0.802	0.278	0.802	0.278	**0.99**	**0.444**	0.762	0.278
KNNeighbors	0.99	0	0.733	0.278	0.733	0.278	0.941	0.167	0.713	0.333
AdaBoost	1	0	0.941	0.278	0.941	0.278	1	0	0.911	0.278

Fonte: Elaborada pelos autores.

É digno de nota que o modelo com melhor desempenho na identificação da classe minoritária é um tão simples como *Naive-Bayes* aliado a Tomek links. Esse resultado, no entanto, encontra sustentação na literatura: em Neo *et al.* (2023), é ele o modelo com melhor *f1-score*, e Roy e Garg (2017), que usam o mesmo conjunto de dados que este trabalho, também elegem *Naive-Bayes* como o que melhor performa; Swana, Doorsamy e Bokoro (2022), ainda que no contexto de *time series*, corroboram o impacto das técnicas de *undersampling* no *Naive-Bayes*, que resulta em performance competitiva em relação a outros modelos.

Em resumo, munido do conjunto de dados com as variáveis *G1* e *G2*, cabe dizer que as propostas de amostragem melhoraram o desempenho dos modelos, com a combinação de *GradientBoosting* e SMOTE alcançando 97.5% de *precision*, identificando 94.4% dos alunos em risco de reprovação no conjunto de testes. Esse resultado de *recall* supera os outros trabalhos citados na Tabela 1 que usaram o mesmo conjunto de dados, e quase todos em *accuracy*, ficando ligeiramente atrás da proposta de Muruganandam e Rajini (2021). Usando o modelo sem as variáveis com maior correlação, as técnicas de amostragem se mostram ainda mais eficientes na identificação da classe minoritária, com *Naive-Bayes* e Tomek links alcançando *accuracy* de 90.8% e identificando quase metade dos alunos em risco de reprovação. Esse resultado não pôde ser comparado com os trabalhos que usam o mesmo conjunto de dados porque é uma análise que os autores não fazem. Finalmente, cabe pontuar que a maioria dos modelos de *machine learning* não foi capaz de aprender nada quando usando o conjunto de dados menos informativo e sem alguma proposta de amostragem, predizendo apenas uma das classes para todos os casos de teste.

4. Considerações Finais

Os principais resultados sugerem que as técnicas de amostragem, especialmente o SMOTE e o Tomek links, desempenham um papel importante na melhora da performance dos modelos de predição de alunos em risco de reprovação, particularmente na

identificação da classe minoritária. Entre os modelos avaliados, o *GradientBoosting*, combinado com o SMOTE, obteve o melhor desempenho geral, alcançando melhorias significativas no *recall* da classe minoritária e performando melhor que os outros trabalhos que usaram o mesmo conjunto de dados. Com um conjunto sem as variáveis relacionadas às notas anteriores (*G1* e *G2*), o modelo *Naive-Bayes*, combinado com Tomek links, se destacou pela performance mais equilibrada entre as classes, apontando que modelos simples, aliados a técnicas de amostragem, podem competir com outros modelos mais complexos e computacionalmente mais caros em cenários de desequilíbrio de classes.

Com um *dataset* com *features* bem correlacionadas com a variável *target*, as técnicas de amostragem se mostraram relevantes na identificação da classe minoritária, aumentando até dez pontos percentuais o *recall* desta classe. No entanto, é com o conjunto de dados sem essas variáveis que a importância das técnicas é evidenciada, fazendo com que modelos que não conseguiam identificar nenhuma instância da classe minoritária passassem a identificar até 0.554. Interessantemente, a combinação de modelo-técnica que melhor performa é de uma de *undersampling* com modelos simples de inteligência máquina, sugerindo que, neste problema específico pelo menos, é mais importante para os modelos a definição clara dos *clusters* no espaço de dados do que mais observações geradas sinteticamente - esse resultado pode ser investigado por trabalhos futuros.

Acerca das técnicas de amostragem, é importante pontuar que SMOTE se destacou com um conjunto de dados mais informativo, alcançando as melhores métricas com modelos baseados em *boosting*, como *GradientBoosting* e *AdaBoost*. Tomek links, por outro lado, se destacou com o conjunto menos informativo, com os modelos *LogisticRegression* e, principalmente, *Naive-Bayes*. A técnica SMOTE + Tomek link performou sempre igual ou pior que SMOTE e a técnica ADASYN performou pior que as outras em ambos os conjuntos de dados, embora ainda tenha se mostrado melhor do que não usar técnica alguma.

Trabalhos futuros podem se debruçar sobre um passo propositalmente negligenciado aqui: *feature engineering* e *feature selection*, isto é, construir novas variáveis e selecionar as variáveis mais correlacionadas com a variável *target.* Esse passo deve melhorar o desempenho global dos modelos e não foi executado aqui para reservar as mudanças no desempenho à, exclusivamente, a implementação das técnicas de amostragem. A eficácia das combinações de modelos-técnicas discutidas também pode ser utilizada em outros contextos além da predição de performance de estudantes, em especial a combinação de Tomek links e modelos mais simples, como o *Naive-Bayes*.

Os resultados de aplicações como as discutidas aqui podem ajudar a mitigar a evasão e reprovação de estudantes em diversos contextos, antecipando alunos em risco e orientando a tomada de decisão sobre medidas corretivas. Cabe pontuar, no entanto, que soluções baseadas em dados demográficos oferecem pouca informação sobre quais as medidas de correção necessárias, não acompanham o desenvolvimento dos estudantes e tendem a ser vulneráveis a problemas de privacidade.

Referências Bibliográficas

Alharbi, I. A.; Almalki, A. J.; Zou, C. C. Hyperparameter Optimization and Comparison of Student Performance Prediction Algorithms. International Conference on Computational Science and Computational Intelligence, 2021.

Ali, D. A.; Aborizk, M.; Dahroug, A. Prediction of students performance by using machine learning techniques. In: 4th International Conference on Artificial Intelligence, Robotics and Control, 2023.

Assistant, M. K. *et al.* Predictive Model for Students' Academic Performance Using Classification and Feature Selection Techniques. In: 2nd International Conference on Computational Methods in Science & Technology, 2021.
Batista, G.; Bazzan, M.; Monard, M. Balancing Training Data for Automated Annotation of Keywords: a Case Study. In: WOB, 10-18, 2003.
Bujang, S. D. A. *et al.* Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review. IEEE Access, vol. 11, pp. 1970-1989, 2023.
Chowdhury, A. *et al*. Student performance Prediction using Ensemble Technique. In: 8th International Conference on Advanced Computing and Communication Systems, 2022.
Costa, E. 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, 2017.
Cortez, P.; Silva, A. Using data mining to predict secondary school student performance. Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008.
CHAWLA, N. V. *et al.* SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 321-357, 2002.
He, H. *et al.* ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks, 2008.
Imran, M. *et al.* Student Academic Performance Prediction using Supervised Learning Techniques. International Journal of Emerging Technologies in Learning, 14, 2019.
Kiu, C. C. Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities. In: Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Malaysia, 2018.
Muruganandam, S.; Rajini, A. An Effective Utilization of Optimal Deep Learning Model Based Student Performance Prediction. In: International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation, 2021.
Neo, A. V. B. S. *et al.* Previsão de reprovação de estudantes utilizando aprendizado de máquina. Neuvas Ideas en Informática Educativa, v. 17, 2023.
Prodanov, C. C.; Freitas, E. C. Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico. 2ed. Novo Hamburgo: Feevale, 2013.
Ram, D. S. *et al.* Machine Learning based student academic performance prediction. In: Proceedings of the Third International Conference on Inventive Research in Computing Applications, 2021.
Razak, N. H. *et al.* Prediction of secondary students performance: a case study. 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2021
Roy, S.; Garg, A. Predicting Academic Performance of Student Using Classification Techniques. In: 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics, 2017.
Santana, M. L.; Gusmão, R. P. De; Gusmão, C. S. D. Predição de Desempenho em Língua Portuguesa de Estudantes do Ensino Fundamental: Um Estudo de Caso em Sergipe. Revista Novas Tecnologias na Educação, Porto Alegre, v. 21, n. 2, p. 198–207, 2023. DOI: 10.22456/1679-1916.137740.
Stoll, B. B.; Cury, D.; Menezes, C. S. Framework para predições e recomendações em dados acadêmicos. Revista Novas Tecnologias na Educação, Porto Alegre, v. 16, n. 2, p. 413–422, 2018. DOI: 10.22456/1679-1916.89244.
Swana, E. F.; Doorsamy, W. Bokoro, P. Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. Sensors, 2022.