DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

IMPACTO DE TÉCNICAS DE AMOSTRAGEM NA PERFORMANCE DE MODELOS DE PREDIÇÃO DE DESEMPENHO DE ALUNOS EM PENSAMENTO COMPUTACIONAL

Carmélia Teixeira de Sousa[i] 0000-0002-4643-2541
Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte
Júlio César da Silva Dantas[ii] 0000-0002-3729-9662
Universidade Federal do Rio Grande do Norte

RESUMO: A crescente evasão e reprovação em cursos de Tecnologia tem encontrado na inteligência artificial uma aliada para mitigar seus impactos. Este estudo investigou a capacidade preditiva de dados demográficos na identificação de alunos em risco de reprovação em turmas de pensamento computacional. Para isso, construiu modelos de *machine learning* e avaliou seus desempenhos, também considerando o impacto de diferentes técnicas de amostragem. Os resultados indicaram uma baixa correlação entre dados demográficos e o desempenho dos alunos, sugerindo que esses dados, isoladamente, possuem limitações preditivas. Contudo, o uso de técnicas como SMOTE mostrou-se significativo, elevando as métricas de avaliação dos modelos. Destacam-se a regressão logística, com *recall* de 53% para a classe de reprovação, e o SVC, que atingiu 75% de acurácia quando combinado com SMOTE e *Tomek links*. Esses achados reforçam o papel de estratégias de balanceamento para melhorar a eficácia preditiva em cenários desafiadores.

PALAVRAS-CHAVE: Predição de performance de estudantes. *Machine Learning*. Pensamento Computacional.

IMPACT OF SAMPLING TECHNIQUES ON THE PERFORMANCE OF PREDICTION MODELS FOR STUDENT PERFORMANCE IN COMPUTATIONAL THINKING

ABSTRACT: The increasing dropout and failure rates in Technology courses have found an ally in artificial intelligence to mitigate their impacts. This study investigated the predictive capacity of demographic data in identifying students at risk of failure in computational thinking classes. For this purpose, machine learning models were developed and their performance evaluated, also considering the impact of different sampling techniques. The results indicated a low correlation between demographic data and student performance, suggesting that these data, in isolation, have limited predictive power. However, the use of techniques such as SMOTE proved significant, enhancing the models' evaluation metrics. Highlights include logistic regression, with a recall of 53% for the failure class, and SVC, which achieved 75% accuracy when combined with SMOTE and Tomek links. These findings emphasize the role of balancing strategies in improving predictive effectiveness in challenging scenarios.

KEYWORDS: Student Performance Prediction. Machine Learning. Computational Thinking.

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

1 Apresentação

Os cursos de graduação das áreas de Ciências Exatas e Aplicadas sofrem, desde muito tempo, com problemas de evasão e reprovação; em geral, um a cada dois alunos não conclui o curso (Souza; Morais; Silva Junior, 2015 e Silva, Silva, Albuquerque, 2016). Diversas ações vêm sendo implementadas a fim de mitigar os impactos desses problemas, a que cabe citar: o investimento em estrutura, revisão e reestruturação de matrizes curriculares e a ampliação de programas de bolsas de pesquisa e inovação (Filho; Siqueira; Leal, 2020).

Um conjunto de soluções que têm ganhado atenção neste contexto nos últimos anos é o uso de algoritmos de *Machine Learning* (ML), criando um novo campo de estudos: *Learning analytics*, que investiga as intersecções entre processos educacionais e estatística. Ainda assim, os estudos envolvendo análise de dados educacionais para a melhoria do desempenho acadêmico são poucos (Guimarães *et al*., 2020).

Com a translação do foco das pesquisas para os ambientes virtuais de aprendizagem, muito associada com a disseminação da modalidade de educação a distância e a facilidade na coleta de dados nesse formato de ambiente, pouca pesquisa no Brasil tem se debruçado na investigação da predição de performance de estudantes em cursos presenciais. Tendo isso dito, este trabalho integra um esforço de desenvolvimento de modelos de *machine learning*, que volta os olhos para classes presenciais e que se baseia, pelo menos em um primeiro passo, em dados demográficos dos alunos; futuramente, deve buscar incorporar outras fontes de dados no processo de modelagem, de maneira a apreender mais dinamicamente o desenvolver do estudante.

Este trabalho tem sua reserva de importância enquanto investiga a eficácia de modelos de ML, baseados em dados demográficos, na predição de alunos em risco de reprovação no componente curricular de Pensamento Computacional, o que poderia ajudar a identificar e dar suporte a esses estudantes, diminuindo os índices de reprovação e evasão dos cursos associados à tecnologia da informação. O trabalho se distancia da literatura em dois pontos especialmente: porque prevê o desempenho no componente de Pensamento Computacional, que é uma das primeiras disciplinas cursadas pelos estudantes, não contando com dados relativos aos históricos dos estudantes em outras disciplinas, como em outros trabalhos; e endereça o problema do desbalanceamento em classes (aprovado/reprovado) com três diferentes abordagens de amostragem. Como questão de pesquisa, tem: modelos clássicos de

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

ML conseguem prever alunos em risco com precisão adequada? E qual o impacto no desempenho dos modelos do uso das técnicas de amostragem?

Para isso, o trabalho coleta, trata e analisa os dados disponíveis no portal de Dados Abertos da UFRN, transforma, treina os modelos, com e sem a aplicação da amostragem, e avalia as principais métricas associadas aos problemas de classificação. Por fim, compara as métricas alcançadas com outras investigações na literatura.

As seções seguintes do artigo são divididas como se segue: primeiro, é abordado o referencial teórico que dá base ao trabalho desenvolvido aqui, em especial no contexto brasileiro; depois é feito o desenho experimental, pontuando os passos e decisões tomadas acerca dos dados e treinamento e avaliação dos modelos; as métricas obtidas são comparadas com a literatura e, finalmente, tem-se as considerações finais e apontamentos para trabalhos futuros.

2 Referencial teórico

2.1 Dados usados

Na literatura, a coleta de dados pode ser dividida em três formas principais: aulas *online*, *offline* e *blended*. Na primeira, os dados são coletados principalmente de um sistema de gerenciamento de aprendizagem (LMS) e consistem em fluxo de cliques, tempo dedicado à aprendizagem, conteúdo postado em fóruns, engajamento em discussões, notas em avaliações e outros recursos relacionados às atividades *online*; Rohani (2023), por exemplo, usa *click-stream* associado a assistir a uma videoaula até o final, pular para frente ou para trás em uma videoaula, pausar, alterar a taxa de reprodução de um vídeo, interagir com um fórum de discussão ou postar uma pergunta no fórum e interagir com um teste, apenas para citar alguns.

As aulas *offline* focam nos registros de aprendizagem dos alunos e informações de contexto social, como educação dos pais, renda familiar, registro familiar e avaliação de desempenho do aluno (Zhang, 2021); além de outros recursos, usa gênero, educação e ocupação dos pais, informações geográficas, informações escolares e se itens diferentes, como um carro ou computador, estavam disponíveis (Cohausz, 2023).

Vale dizer que há trabalhos que divergem consideravelmente dessas fontes de dados, como em Caruso *et al*. (2022), que para prever a compreensão de leitura do usuário, usa um

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

dispositivo de olhar para rastrear os movimentos dos olhos durante a leitura; Wampfler *et al*. (2019) prevê o estado afetivo do usuário por meio de sinais coletados com dados de biossensor de condutância da pele, medidas cardíacas e temperatura da pele, e Sailema *et al*. (2020), que extrai a emoção (raiva, nojo, medo, felicidade, neutralidade, tristeza, surpresa, desprezo) das gravações dos alunos em aulas online para prever seu desempenho. Neste trabalho, os dados coletados são demográficos, vieram do programa de dados abertos da UFRN e são discutidos adiante.

2.2 O estado de pesquisas no Brasil

O trabalho de Neo *et al*. (2023) investiga o desempenho de seis algoritmos de aprendizagem de máquina na predição do desempenho de estudantes do ensino técnico; coleta dados demográficos como munícipio, raça, renda e número de pessoas na casa de 25 estudantes e tem o *Waikato Environment for Knowledge Analysis* (Weka) como ferramenta para o desenvolvimento dos experimentos computacionais. Os autores chegam a métricas de precisão de 73.1% com árvores de decisão e 64.5% de F1-*Score* com *Naive-Bayes*. Garcia *et al*. (2023) usam exclusivamente o histórico acadêmico de estudantes do curso de Ciências da Computação em 15 disciplinas e alcançam métricas como 72.60% de acurácia e 80% de *recall* para a classe de reprovados.

Evandro *et al*. (2017) se distancia destes trabalhos usando duas fontes de dados: um banco de dados associado à educação à distância e a um curso presencial, contendo informações como acesso à plataforma, renda, estado civil, participação em discussões e notas em avaliações. Os autores chegaram a 82% de F1-*Score* com árvores de decisão, e uma efetividade que varia entre 55%-82% nos cursos à distância e 50% e 79% no curso presencial, a depender do algoritmo usado.

2.3 *Synthetic minority over-sampling technique* (SMOTE)

Quando se trabalha com conjuntos de dados desbalanceados, onde as classes a serem previstas têm distribuições desiguais, duas abordagens comuns no aprendizado de máquina são: atribuir pesos diferentes para cada classe durante o treinamento ou ajustar a amostragem

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

dos dados, fazendo *oversampling* da classe minoritária ou *undersampling* da classe majoritária. A técnica *Synthetic Minority oversampling Technique* (SMOTE), por sua vez, oferece uma solução alternativa, criando novas instâncias sintéticas com base nos dados existentes. A implementação adotada segue a proposta de Chawla *et al*. (2002).

Com o SMOTE, cada exemplo da classe minoritária é analisado e um conjunto de vizinhos mais próximos (geralmente calculado com base na distância Euclidiana) é selecionado. A partir disso, novos exemplos são gerados através da interpolação entre o exemplo original e os seus vizinhos (Chawla *et al*., 2002).

O uso do SMOTE reduz o viés que os modelos de aprendizado de máquina podem apresentar em relação à classe majoritária. Ao gerar instâncias sintéticas, a técnica melhora o desempenho na classificação da classe minoritária. No entanto, é importante destacar que o SMOTE pode ser vulnerável a *outliers*, já que a interpolação entre os pontos pode amplificar a influência de dados atípicos.

2.4 *Oversampling* e *undersampling*: SMOTE e *Tomek Links*

A técnica discutida anteriormente, SMOTE, pode gerar amostras ruidosas interpolando novos pontos entre *outliers*. Frequentemente, os *clusters* de classes não são bem definidos, visto que alguns exemplos de classes majoritárias podem estar invadindo o espaço de classes minoritárias. O oposto também pode ser verdade: a interpolação de exemplos de classes minoritárias pode expandir o *cluster* de classes minoritárias, introduzindo exemplos artificiais de classes minoritárias no espaço de classes majoritárias. Ter um classificador nesta situação pode levar ao *overfitting*, como uma árvore de decisão que pode ter que criar vários ramos para distinguir entre os exemplos que estão no lado errado da borda de decisão (Batista, Bazzan, Monard, 2003).

Sendo assim, a técnica de SMOTE pode ser combinada a outra técnica de amostragem, essa de *undersampling*, como a *Tomek links*. Um *link* de *Tomek* pode ser definido da seguinte forma: 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 *link* de *Tomek* 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). Se dois exemplos formam um *link* de *Tomek*, então um desses exemplos é ruído ou ambos os exemplos estão na fronteira das classes. Essa técnica é usada frequentemente para identificar e remover

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

exemplos que causam confusão entre as classes em problemas de classificação. A implementação da técnica aqui se dá como definida por Batista, Bazzan, Monard (2003).

2.5 *Oversample using adaptive synthetic* (ADASYN)

Este método é semelhante ao SMOTE, mas gera um número diferente de amostras dependendo de uma estimativa da distribuição local da classe a sofrer *oversampling*. A ideia central do algoritmo ADASYN é usar uma distribuição de densidade como critério para decidir automaticamente o número de amostras sintéticas que precisam ser geradas para cada exemplo da classe minoritária; é uma medida da distribuição de pesos para diferentes exemplos da classe minoritária de acordo com seu nível de dificuldade de aprendizado. O conjunto de dados resultante após a aplicação do ADASYN não apenas fornecerá uma representação balanceada da distribuição de dados, mas também forçará o algoritmo de aprendizado a se concentrar nos exemplos mais difíceis de aprender. Essa é uma diferença importante em relação ao algoritmo SMOTE, no qual o mesmo número de amostras sintéticas é gerado para cada exemplo da classe minoritária (He *et al*., 2008). A implementação usada aqui é como a definida em He *et al*. (2008).

A principal diferença entre o ADASYN e o SMOTE está na forma como cada algoritmo gera as amostras sintéticas para a classe minoritária. Enquanto o SMOTE gera o mesmo número de amostras sintéticas para cada exemplo da classe minoritária, o ADASYN ajusta dinamicamente o número de amostras sintéticas com base na dificuldade de aprendizado de cada exemplo. No ADASYN, uma distribuição de densidade é usada para determinar quantas amostras precisam ser criadas para cada ponto, focando mais nos exemplos que são mais difíceis de aprender. Isso faz com que o ADASYN não apenas equilibre a distribuição dos dados, mas também force o algoritmo de aprendizado a dar mais atenção aos exemplos mais complicados. Em contraste, o SMOTE trata todos os exemplos da classe minoritária de forma igual, gerando uma quantidade fixa de amostras para cada um.

3 Material e métodos

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

Esta pesquisa é um esforço introdutório de mobilizar conhecimentos teóricos em uma aplicação prática, na resolução de determinado problema; por isso, pode ser classificada como exploratória e aplicada. Por que realiza um experimento computacional na proposta de solução e por que baseia a análise dos resultados em evidências quantificáveis, assume o paradigma metodológico orientado à pesquisa experimental e quantitativa (Prodanov, Freitas, 2013). O conjunto de dados foi coletado a partir do conjunto de dados abertos da UFRN, possui 227 observações e são descritos no quadro abaixo:

Tabela 1 - Variáveis usadas.

Variável	Espaço	Descrição
media_final	0,1	Classificação aprovado/reprovado.
Sexo	0,1	Sexo declarado do estudante.
Raca	0-4	Raça declarada pelo estudante.
estado_origem	0,1	Identifica se o aluno é do RN ou não.
n_subjects	1-8	Número de disciplinas cursadas no semestre.
Renda	0-30000	Renda declarada pelo estudante.
escola_ens_medio	0-4	Tipo de escola em que o estudante cursou o ensino médio.
auxilio_alimentacao	0,1	Se o estudante possui auxílio alimentação no semestre.
auxilio_moradia	0,1	Se o estudante possui auxílio moradia no semestre.
Idade	20-58	Idade do estudante no semestre.

Fonte: Elaborado pelos autores, 2024.

Os dados foram extraídos da disciplina de Pensamento Computacional oferecida pelo Instituto Metrópole Digital entre os anos de 2018 e 2022. Observações com elementos faltando foram removidos. Os experimentos foram realizados usando as bibliotecas *numpy* (2.1.1), *pandas* (2.2.2), *scikit-learn* (1.5.1) e *xgboost* (2.1.1). Todos os modelos de *machine learning* 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*; as variáveis n_subjects, renda, idade passaram por *scaling* (MinMaxScaler) e o conjunto por *sampling* (SMOTE). As variáveis raca e escola_ens_medio sofreram *encoding* em colunas. Nenhuma das variáveis têm correlação especialmente relevante com a variável *target*. As variáveis binárias tiveram sua correlação investigada através do chi-teste de independência e o coeficiente de *Chamér*.

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

3 Resultados

Os desempenhos dos modelos de inteligência de máquina estão descritos no quadro abaixo e uma análise mais cuidadosa dos modelos é feita adiante, pontuando, em especial, o impacto das técnicas de amostragem.

Tabela 3 – Métricas de avaliação dos modelos de machine learning

Modelo	Scaler				Scaler + ADASYN				Scaler + SMOTETomek				Scaler + SMOTE			
	acc	pre	rec	f1	acc	pre	rec	f1	acc	Pre	rec	f1	acc	pre	Rec	f1
RandomForest	0.68	0.68	0.68	0.68	0.68	0.65	0.68	0.68	0.67	0.66	0.67	0.66	0.67	0.65	0.67	0.66
GradientBoosting	0.71	0.67	0.71	0.69	0.62	0.62	0.62	0.62	0.65	0.65	0.65	0.65	0.72	0.71	0.72	0.72
MLPClassifier	0.67	0.66	0.67	0.66	0.70	0.69	0.70	0.69	0.71	0.70	0.71	0.70	0.70	0.70	0.70	0.70
LogisticRegression	0.75	0.57	0.75	0.65	0.67	0.71	0.67	0.68	0.70	0.71	0.70	0.70	0.71	0.73	0.71	0.72
SVC	0.74	0.69	0.74	0.7	0.59	0.61	0.59	0.6	0.75	0.75	0.75	0.75	0.74	0.74	0.74	0.74
XGBoost	0.71	0.67	0.71	0.69	0.64	0.64	0.64	0.64	0.64	0.66	0.64	0.65	0.65	0.65	0.65	0.65
Naive-bayes	0.77	0.75	0.77	0.7	0.52	0.66	0.52	0.55	0.49	0.62	0.49	0.53	0.49	0.62	0.49	0.53
KNeighbors	0.72	0.69	0.72	0.7	0.61	0.64	0.61	0.62	0.65	0.65	0.65	0.65	0.64	0.64	0.64	0.64
AdaBoost	0.75	0.57	0.75	0.65	0.57	0.61	0.57	0.58	0.65	0.69	0.65	0.67	0.64	0.69	0.64	0.66

Fonte: Elaborado pelos autores, 2024.

Discutindo primeiro os resultados sem amostragem, faz parecer que aqueles mais simples, como *naive-bayes* e *logistic regression*, foram os que melhor performaram. No entanto, como denuncia as métricas de *recall* na tabela 2, os modelos não conseguiram diferenciar bem as classes: *naive-bayes*, por exemplo, entre todos os estudantes previstos com risco de reprovação, só identifica corretamente 12% deles e alcança um valor de acurácia de 77% porque prevê a classe majoritária para a maioria dos casos de teste. *Logistic regression* é ainda pior: o modelo não prevê corretamente nenhum estudante em risco, alcançando *precision*, *recall* e f1-*score* igual a zero para a classe de alunos em risco de reprovação; *Adaboost* também. O modelo que tem o melhor desempenho de *recall* para a classe de alunos

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

em risco de reprovação, isto é, a porcentagem de acertos do modelo em razão do número observações da classe, foi o modelo *RandomForest*, que conseguiu identificar 35% dos alunos em risco de reprovação durante o teste.

Tabela 4 - Resultados de recall para cada uma das classes

Modelos	**Base**		**SMOTE**		**SMOTE + Tk**		**ADASYN**	
	apr	**rep**	**apr**	**rep**	**apr**	**rep**	**apr**	**rep**
SVC	0.92	0.18	0.83	0.47	0.85	0.47	0.71	0.24
RandomForest	0.79	0.35	0.81	0.24	0.79	0.29	0.83	0.24
GradientBoosting	0.87	0.24	0.85	0.35	0.77	0.29	0.75	0.24
MLPClassifier	0.79	0.29	0.79	0.41	0.83	0.35	0.81	0.35
LogisticRegression	1	0	0.77	**0.53**	0.77	0.47	0.71	0.53
XGBoost	0.87	0.24	0.77	0.29	0.73	0.35	0.75	0.29
Naive-bayes	0.98	0.12	0.5	0.47	0.5	0.47	0.5	**0.59**
KNNeighbors	0.88	0.24	0.75	0.29	0.77	0.29	0.69	0.35
AdaBoost	1	0	0.69	0.47	0.71	0.47	0.65	0.29

Fonte: Elaborado pelos autores, 2024.

Usando a *Synthetic Minority Over-sampling Technique* para sintetizar mais dados da classe minoritária, alunos que reprovaram, o desempenho dos modelos muda consideravelmente. Entre os modelos, todos, exceto *Random Forest*, aumentaram a performance na identificação de estudantes em risco de reprovação. Cabe destacar que *Logistic regression* é um modelo que performa consideravelmente melhor com a técnica de amostragem SMOTE, alcançado 53% de *recall* na classe de alunos em risco e 77% em alunos aprovados. SVC e MLPClassifier também melhoraram com a amostragem em métricas gerais e na identificação de alunos em risco.

Unindo a técnica de *Tomek links* a SMOTE, os modelos *Adaboost*, SVC, *Logistic regression* e *Naive-bayes* empatam em avaliação, conseguindo identificar apenas 47% dos alunos em risco no conjunto de teste, apresentando queda na performance em relação ao modelo que usava apenas SMOTE, mas ainda performando melhor do que sem a técnica de

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

amostragem. Finalmente, com *Oversample Using Adaptive Synthetic*, os modelos têm performance levemente piorada, em relação ao SMOTE.

4 Conclusão

Um dos aspectos mais desafiadores revelados pela pesquisa foi a baixa correlação entre os atributos demográficos e o desempenho acadêmico dos alunos. Isso indica que os dados demográficos, isoladamente, podem não ser suficientes para construir modelos robustos de predição de reprovação. Portanto, futuras pesquisas devem incorporar outras fontes de dados, como o histórico acadêmico, engajamento em plataformas de aprendizagem, e até aspectos comportamentais, para aumentar a acurácia dos modelos e fornecer informações mais detalhadas sobre os fatores que levam ao sucesso ou fracasso dos estudantes.

Ademais, ao comparar diferentes técnicas de amostragem, foi observado que, embora o SMOTE tenha gerado bons resultados, sua combinação com *Tomek Links* não trouxe melhorias adicionais significativas, e em alguns casos, reduziu o desempenho. Isso sugere que a remoção de amostras ruidosas nem sempre é benéfica e deve ser aplicada com cautela, especialmente em conjuntos de dados menores. O ADASYN, que prioriza a criação de amostras sintéticas para exemplos mais difíceis, também não apresentou ganhos consideráveis em comparação com o SMOTE, reforçando que a simplicidade do SMOTE pode ser mais eficaz para esse tipo de cenário educacional.

Do ponto de vista prático, este trabalho reforça a viabilidade do uso de *machine learning* para a predição de risco de reprovação, mas também destaca a necessidade de cuidados ao tratar de problemas de desbalanceamento de classes. Embora as técnicas de *oversampling* melhorem a identificação de alunos em risco, o contexto dos dados e as características do conjunto de treinamento devem ser sempre considerados ao escolher os métodos de amostragem e os algoritmos de predição.

Como desdobramento para estudos futuros, sugere-se a inclusão de variáveis que vão além dos dados demográficos, como o desempenho em atividades realizadas ao longo do curso e informações extraídas de plataformas de ensino, o que permitirá uma análise mais completa e precisa. Além disso, a ampliação da base de dados e a utilização de abordagens mais sofisticadas, como redes neurais ou modelos temporais, podem proporcionar *insights* mais profundos e aumentar a capacidade preditiva das soluções.

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

Referencias

AMRA, I. A. A.; MAGHARI, Y. A. A. Students Performance Prediction Using KNN and Naïve Bayesian. *In*: International Conference on Information Technology, 8, 2017, Amã, Jordânia. **Anais eletrônicos** [...] Amã: IEEE, 2017. p. 909-913, doi: 10.1109/ICITECH.2017.8079967. Disponível em: https://ieeexplore.ieee.org/document/8079967 Acesso em: 12 nov. 2024.

BATISTA, G.; BAZZAN, B.; MONARD, M. Balancing Training Data for Automated Annotation of Keywords: a Case Study. *In*: **WOB**, p. 10-18, 2003. Disponível em: https://www.inf.ufrgs.br/maslab/masbio/papers/balancing-training-data-for.pdf Acesso em: 12 nov. 2024.

CARUSO, M.; PEACOCK, C. E.; SOUTHWELL, R.; ZHOU, G.; D'MELLO, S. K. Going deep and far: Gaze-based models predict multiple depths of comprehension during and one week following reading. *In*: **Proceedings of the 15th International Conference on Educational Data Mining**, 2022. Disponível em: https://educationaldatamining.org/edm2022/proceedings/2022.EDM-long-papers.13/2022.EDM-long-papers.13.pdf Acesso em: 12 nov. 2024.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, 321-357, 2002. Disponível em: https://arxiv.org/abs/1106.1813 Acesso em: 11 nov. 2024.

COHAUSZ, L.TSCHALZAV, A.; BARTELT, C. Investigating the importance of demographic features for edm-predictions. *In*: **Proceedings of the 16th International Conference on Educational Data Mining**, 2023. Disponível em: https://educationaldatamining.org/EDM2023/proceedings/2023.EDM-long-papers.11/index.html Acesso em 10 nov. 2024.

COSTA, E. B.; FONSECA, B.; SANTANA, M. A.; ARAÚJO, F. F.; REGO, J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. **Computers in Human Behavior**, v. 73, p. 247-256, 2017. Disponível em: https://www.infona.pl/resource/bwmeta1.element.elsevier-7608d17d-609f-3558-847a-38d53476cec1 Acesso em: 05 out. 2024.

GARCIA, L. M. L. S.; LARA, D. S.; GOMES, R. S.; CAZELLA, S. C. Ferramenta para Predição do Desempenho Acadêmico no Ensino Superior. *In*: Simpósio Brasileiro de Informática na Educação (SBIE), **Anais eletrônicos** [...], 34. 2023, Passo Fundo/RS. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 1215-1225. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/26748 Acesso em: 03 out. 2024.

GOTTARDO, E.; KAESTNER, C. A. A.; NORONHA, R. V. Estimativa de Desempenho Acadêmico de Estudantes: Análise da Aplicação de Técnicas de Mineração de Dados em Cursos à Distância. **Revista Brasileira de Informática na Educação**, 22, 01, 2014. Disponível em: http://milanesa.ime.usp.br/rbie/index.php/rbie/article/view/2381 Acesso em: 12 out. 2024.

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

GUIMARÃES, C. A.; NUNES, I.; PIRES, A. K.; ALENCAR, E. A Produção de Learning Analytics e Predição de Desempenho Acadêmico por pesquisadores Brasileiros: Uma Revisão Sistemática da Literatura. *In*: V Congresso sobre Tecnologias na Educação, **Anais eletrônicos** [...]. 2020. Disponível em: https://sol.sbc.org.br/index.php/ctrle/article/view/11408 Acesso em: 23 nov. 2024.

FILHO, F. H. B.; SIQUEIRA, D. S.; LEAL, B. C. Predição de Evasão Utilizando Técnicas de Classificação: Um Estudo de Caso do Instituto Federal do Ceará. *In*: Escola Regional de Computação do Ceará, Maranhão e Piauí (ERCEMAPI), **Anais eletrônicos** [...], 8, 2020, Evento Online. Porto Alegre: Sociedade Brasileira de Computação, 2020. Disponível em: https://sol.sbc.org.br/index.php/ercemapi/article/view/11478/11341 Acesso em: 05 nov. 2024.

HÄMÄLÄINEN, W.; VINNI, M. Classifiers for Educational Data Mining. *In*: **Romero *et al*. Handbook of Educational Data Mining**. Flórida, CRC Press, p. 57-71, 2011. Disponível em: https://www.researchgate.net/publication/260300354_Classiers_for_educational_data_mining Acesso em: 05 nov. 2024.

HE, H.; BAI, Y.; GARCIA, E. A.; ADASY, N. Adaptive synthetic sampling approach for imbalanced learning. *In*: **IEEE International Joint Conference on Neural Networks**, 2008. Disponível em: https://ieeexplore.ieee.org/document/4633969 Acesso em 23 nov. 2024.

HOQ, M.; BRUSILOVSKY, P.; AKRAM, B. Analysis of an Explainable Student Performance Prediction Model in an Introductory Programming Course. *In:* Educational Data Mining Conference, **Anais eletrônicos** [...], 2023. Disponível em: https://files.eric.ed.gov/fulltext/ED630852.pdf Acesso em 05 nov. 2024.

JYOTI, E.; WALIA, E. A. S. A review on recommendation system and web usage data mining using k-nearest neighbor (knn) method. **Int. Res. J. Eng. Technol**. 4, 2017. Disponível em: https://www.irjet.net/archives/V4/i4/IRJET-V4I4713.pdf Acesso em: 08 nov. 2024.

NEO, A.; NEO, G. S.; FREITAS JR, O. G.; RODRIGUES, W. R. M. Previsão de reprovação de estudantes utilizando Aprendizado de Máquina. IN: Sánchez, J. (2023) Editor. **Nuevas Ideas en Informática Educativa**, Vol. 17, p. 413 – 417, Santiago de Chile. Disponível em: https://www.researchgate.net/publication/376453197_Previsao_de_reprovacao_de_estudantes_utilizando_Aprendizagem_de_Maquina Acesso em 12 nov. 2024.

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho acadêmico. 2ed. Novo Hamburgo: Feevale, 2013. Disponível em: https://www.feevale.br/Comum/midias/0163c988-1f5d-496f-b118-a6e009a7a2f9/E-book%20Metodologia%20do%20Trabalho%20Cientifico.pdf Acesso em 05 out. 2024.

ROHANI , N.; GAL , K.; GALLAGHER , M.; MANATAKI, A. Early prediction of student performance in a health data science MOOC . In: FENG, M., KÄSER, T.; TALUKDAR P., *In:* Proceedings of the 16th international conference on educational data mining, **Anais eletrônicos** [...], 2023. Disponível em: https://educationaldatamining.org/EDM2023/proceedings/2023.EDM-short-papers.32/2023.EDM-short-papers.32.pdf Acesso em: 15 out. 2024.

SAA, A. A. Educational data mining & students performance prediction. **Int. J. Adv. Comput**. Sci. Appl. 7, 212–220, 2016. Disponível em: https://www.researchgate.net/publication/303869038_Educational_Data_Mining_Students'_Performance_Prediction Acesso em: 11 nov. 2024.

SAILEMA, W. G. C.; SÁNCHEZ, M.; COREZO, R.; ROMERO, C. Predicting students' performance using emotion detection from face-recording video when interacting with an ITS" *In*: Proceedings of The 13th International Conference on Educational Data Mining, **Anais eletrônicos** [...], 2020. Disponível em: https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_232.pdf Acesso em: 23 out. 2024.

SILVA, E. E. O.; SILVA, J. S.; ALBUQUERQUE, C. H. Uma Análise da Evasão Escolar nos Cursos de Tecnologia da Informação: Um estudo de caso em Floresta/PE. *In*: Workshop sobre Educação em Computação (WEI), **Anais eletrônicos [...]**, 24, 2016, Porto Alegre. Disponível em: https://sol.sbc.org.br/index.php/wei/article/view/9685 Acesso em: 21 out. 2024.

SOUZA, O.; MORAIS, P.; SILVA JÚNIOR, F. Um Estudo sobre a Evasão no Curso de Licenciatura em Informática do IFRN – Campus Natal – Zona Norte. *In*: Workshop sobre Educação em Computação (WEI), 23, 2015, Recife. **Anais eletrônicos [...]**. Porto Alegre: Sociedade Brasileira de Computação, 2015, p. 216-225. DOI: https://doi.org/10.5753/wei.2015.10238. Disponível em: https://sol.sbc.org.br/index.php/wei/article/view/10238 Acesso em: 09 nov.2024.

WAMPFLER, R.; KLINGLER, S.; SOLENTHALER, B.; SCHINAZI, V. R. Affective State Prediction in a Mobile Setting using Wearable Biometric Sensors and Stylus. *In*: Proceedings of The 12th International Conference on Educational Data Mining, **Anais eletrônicos** [...], 2019. Disponível em: https://files.eric.ed.gov/fulltext/ED599181.pdf Acesso em: 10 out. 2024

ZHANG, Y.; YUN, Y.; A. N., R.; CUI, J.; DAI, H.; SHANG, X. Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. **Frontiers in Psychology**. DOI: 10.3389/fpsyg.2021.698490. Disponível em: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.698490/full Acesso em: 23 nov. 2024.

Recebido:	**10/10/2024**
Publicado:	**15/12/2024**

DOI 10.33872/revcontrad.v5n2.e067 | v. 5, n. 2 | Jul./Dez. 2024

[i] Carmélia Teixeira de Sousa é Mestra em Filosofia pela Universidade Federal do Rio Grande do Norte e professora do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte - IFRN. E-mail: carmeliasousa51@gmail.com

[ii] Júlio César da S. Dantas é licenciado em Física (IFRN), mestre em Inovação em Tecnologias Educacionais (UFRN) e aluno do Programa de Pós-graduação em Sistemas e Computação (PPgSC), UFRN. E-mail: dantas.j.c.s@gmail.com