Previsão do Desempenho Acadêmico de Estudantes do Ensino Superior Utilizando Técnicas de Aprendizado de Máquina

Alan Marques da Rocha[1]*; Caroline Belisário Zorzal[2]

[1] Universidade Federal do Ceará. Mestrando em Engenharia Elétrica e Computação. Programa de Pós-Graduação em Engenharia Elétrica e Computação, Rua Cel. Estanislau Frota; 62010-560; Sobral, CE, Brasil.
[2] Universidade Federal do Mato Grosso. Professora assistente na Faculdade de Engenharia - FAENG/UFMT. Av. Fernando Corrêa da Costa, n° 2367, Bairro Boa Esperança; 78060-900; Cuiabá, MT, Brasil.
*autor correspondente: alanmarquesrocha@usp.br

Previsão do Desempenho Acadêmico de Estudantes do Ensino Superior Utilizando Técnicas de Aprendizado de Máquina

Resumo

As instituições de ensino superior coletam uma grande quantidade de dados sobre o desempenho dos seus alunos, tornando-se um campo fértil para a geração de “insights” por meio da aplicação de algoritmos de aprendizado de máquina. Este trabalho propõe realizar uma análise de predição do desempenho acadêmico dos alunos de instituições de ensino superior utilizando técnicas de aprendizado de máquina. Para realizar essa tarefa, os algoritmos Rede Neural Artificial do tipo Perceptron Multicamadas, Floresta Aleatória e Árvore de Decisão foram implementados. Dois experimentos de classificação foram realizados para cada algoritmo. No primeiro, a técnica SMOTE foi empregada para lidar com o desbalanceamento das classes presentes na base de dados antes do processo de treinamento, teste e validação dos modelos. No segundo experimento, o método de validação cruzada estratificada foi utilizado com o conjunto de dados desbalanceado. A principal contribuição do trabalho é fornecer uma análise comparativa dos algoritmos para a resolução da problemática, além de fornecer uma ferramenta de baixo custo computacional para lidar com desafios comuns na previsão do sucesso acadêmico dos alunos de curso superior em diversas instituições de ensino. O algoritmo Floresta Aleatória obteve uma acurácia de 80,68% no conjunto de teste utilizando a técnica SMOTE, destacando-se dentre os demais algoritmos propostos neste trabalho.
Palavras-chave: Performance acadêmica, Classificação multiclasse, Redes neurais artificiais, Árvore de decisão, Floresta aleatória.

Academic Performance Prediction of Higher Education Students Using Machine Learning Techniques

Abstract

Higher education institutions collect a vast amount of data on their students' performance, becoming a fertile field for generating insights through the application of machine learning algorithms. This work proposes to perform an analysis to predict the academic performance of higher education institution students using machine learning techniques. To accomplish this task, the Artificial Neural Network algorithm of the Multilayer Perceptron type, Random Forest, and Decision Tree were implemented. Two classification experiments were conducted for each algorithm. In the first one, the SMOTE technique was employed to address the class imbalance present in the dataset before the training, testing, and validation processes of the models. In the second experiment, the stratified cross-validation method was used with the unbalanced dataset. The main contribution of this work is to provide a comparative analysis of the algorithms for addressing the problem, as well as to offer a computationally cost-effective tool to tackle common challenges in predicting the academic success of higher education students across various institutions. The Random Forest algorithm achieved an accuracy of 80.68% on the test set using the SMOTE technique, standing out among the other algorithms proposed in this work.
Keywords: Academic performance, Multiclass classification, Artificial neural networks, Decision tree, Random Forest.

Introdução

Instituições de ensino superior em todo o mundo enfrentam os desafios de lidar com diversos estilos de aprendizagem dos alunos com seus respectivos desempenhos

acadêmicos e buscam promover uma experiência de aprendizagem satisfatória e eficiente de forma geral. Para isso, é de interesse das instituições adotar medidas que melhorem o engajamento dos alunos, contribuindo para um melhor desempenho acadêmico dos mesmos.

Muitas instituições procuram adotar medidas para prever e antecipar possíveis dificuldades dos alunos, a fim de construir estratégias de suporte e orientação para aqueles que possam estar em risco de dificuldades acadêmicas ou evasão. Ademais, as instituições coletam anualmente uma grande quantidade de dados sobre o desenvolvimento acadêmico de seus discentes, bem como informações demográficas e socioeconômicas. A combinação desses fatores torna-se um campo fértil para a contribuição de abordagens de Aprendizado de Máquina ("Machine Learning [ML]") na previsão do desempenho acadêmico dos alunos.

O grande volume de dados que são coletados anualmente nas instituições de ensino pode fornecer informações valiosas sobre seus alunos, porém são necessárias técnicas bem elaboradas para realizar uma análise que seja capaz de gerar ferramentas a fim de mitigar os problemas envolvendo seus alunos. Diante disso, técnicas de ML têm sido amplamente utilizadas, tendo em vista o alto nível de precisão em classificações e predições através de algoritmos robustos e de fácil implementação (Mduma et al., 2019). A literatura tem apresentado modelos de predição de sucesso e insucesso acadêmico em cursos de nível superior em várias instituições de ensino utilizando diversas técnicas e abordagens. Muitos trabalhos estão interessados em identificar estudantes que podem ter dificuldades em concluir ou desistir do seu curso.

Hoffait e Schyns (2017) abordaram métodos baseados nos algoritmos de Floresta Aleatória ("Random Forest [RF]"), Regressão Logística ("Logistic Regression [LR]") e Rede Neural Artificial [RNA] para identificar perfis de calouros que apresentaram dificuldades para concluir seu primeiro ano de curso. No trabalho de Beaulac e Rosenthal (2019) foram analisados um grande conjunto de dados de uma universidade no Canadá, contendo informações de 38.842 alunos, para prever o sucesso acadêmico usando apenas o algoritmo RF. Miguéis et al. (2018) implementaram os algoritmos RF e "AdaBoost" para prever o desempenho acadêmico disponível no final do primeiro ano acadêmico na referida base, obtendo uma acurácia média de 96%. Thammasiri e Delen (2013) realizaram um estudo sobre a previsão de desistência de alunos usando diversas técnicas de balanceamento de classes em conjunto com os classificadores RF, RNA, LR e "Support Vector Machine [SVM]". O estudo envolveu um conjunto de dados com 21.654 alunos, onde os autores compararam a eficácia de técnicas de subamostragem aleatória, sobreamostragem aleatória e sobreamostragem sintética. Os melhores resultados foram obtidos com a técnica "Synthetic Minority Over-sampling Technique [SMOTE]" (Chawla e

Bowyer, 2002), que tem sido bem-sucedida na resolução de problemas onde existe um desequilíbrio entre as classes em diferentes domínios.

Tendo em vista que a análise da vida acadêmica dos alunos de cursos de ensino superior é amplamente investigada na literatura em busca de “insights” que possam evitar uma possível evasão, este trabalho propõe um estudo que visa utilizar os algoritmos de ML Rede Neural Artificial do tipo Perceptron Multicamadas [RNA-MLP], RF e Árvore de Decisão (“Decision Tree [DT]”), para realizar a predição do sucesso e insucesso acadêmico dos estudantes de cursos do ensino superior, Tais algoritmos foram escolhidos, considerando sua fácil implementação, baixo custo computacional e por apresentam resultados satisfatórios em problemas de classificação multiclasse em diferentes domínios. Dois experimentos de classificação serão realizados para cada modelo. No primeiro a técnica SMOTE será empregada para lidar com o desbalanceamento das classes presentes na Base de Dados [BD]. No segundo experimento, o método de Validação Cruzada Estratificada (“stratified k-fold”) será utilizado com o conjunto de dados desbalanceado. O objetivo é comparar o desempenho desses algoritmos na tarefa de prever o sucesso e insucesso acadêmico dos alunos, utilizando as técnicas de desbalanceamento e validação cruzada estratificada à fim de gerar “insights” que possam contribuir para medidas e tratativas que reduzam a evasão e baixo nível acadêmico dos alunos.

A principal contribuição do presente trabalho é fornecer uma análise comparativa de diferentes algoritmos de ML na predição do sucesso acadêmico, bem como apresentar abordagens para lidar com desafios comuns encontrados nessa área de pesquisa, como por exemplo o desbalanceamento de classes na BD. Essas contribuições podem ajudar a orientar futuras pesquisas na área da educação e ML, auxiliando na adoção de técnicas mais eficazes para a predição do sucesso dos estudantes.

Material e Métodos

Dados

A BD “Predict Students’ Dropout and Academic Success”[1] foi criada a partir do estudo realizado pelo Instituto Politécnico de Portalegre, Valoriza (2020), possuindo informações sobre matrículas de alunos em cursos de ensino superior de 17 graduações de diferentes áreas do conhecimento, como por exemplo: administração, enfermagem, jornalismo, educação física, arquitetura, etc. Ademais, a BD conta com informações socioeconômicas e macroeconômicas coletadas no momento da matrícula do aluno e no

[1] https://archive-beta.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

final dos primeiro e segundo semestres. A Tabela 1 apresenta a quantidade e os nomes dos atributos, além de fornecer uma informação sobre o tipo do atributo (inteiro, contínuo ou categórico).

Tabela 1. Atributos disponíveis da base de dados.

N°	Nome do Atributo	Tipo
1	Estado civil	Categórica
2	Modo de aplicação	Categórica
3	Ordem de aplicação	Categórica
4	Curso	Categórica
5	Atendimento diurno/noturno	Categórica
6	Qualificação anterior	Categórica
7	Qualificação anterior (grau)	Contínua
8	Nacionalidade	Categórica
9	Qualificação da mãe	Categórica
10	Qualificação do pai	Categórica
11	Ocupação da mãe	Categórica
12	Ocupação do pai	Categórica
13	Nota de admissão	Contínua
14	Estrangeiro	Categórica
15	Necessidades educacionais especiais	Categórica
16	Atraso de mensalidade	Categórica
17	Mensalidades em dia	Categórica
18	Gênero	Categórica
19	Bolsista	Categórica
20	Idade na inscrição	Categórica
21	Internacional	Categórica
22	Unidades curriculares 1º semestre (creditado)	Categórica
23	Unidades curriculares 1º semestre (inscrito)	Categórica
24	Unidades curriculares 1º semestre (avaliações)	Categórica
25	Unidades curriculares 1º semestre (aprovado)	Categórica
26	Unidades curriculares 1° semestre (grau)	Categórica
27	Unidades curriculares 1° semestre (sem avaliações)	Categórica
28	Unidades curriculares 2° semestre (creditado)	Categórica
29	Unidades curriculares 2° semestre (inscritos)	Categórica
30	Unidades curriculares 2° semestre (avaliações)	Categórica
31	Unidades curriculares 2° semestre (aprovado)	Categórica
32	Unidades curriculares 2° semestre (grau)	Categórica
33	Unidades curriculares 2° semestre (sem avaliações)	Categórica
34	Taxa de desemprego	Contínua

35	Taxa de inflação	Contínua
36	PIB	Contínua
-	Classes	Categórica

Fonte: Dados originais da pesquisa.

Os dados estão disponíveis no formato “Comma-separated values [CSV]”. Os atributos que possuem variáveis do tipo categórica foram representados por valores inteiros. Já os atributos do tipo contínuo representam um intervalo de números. Alguns exemplos são apresentados na Tabela 2. As categorias dos demais atributos podem ser consultadas em Valoriza (2020).

Tabela 2. Exemplo de representação dos atributos na base de dados

N°	Nome do Atributo	Tipo	Representação
1	Estado civil	Categórica	1 – solteiro 2 – casado 3 – viúvo 4 – divorciado 5 – união estável 6 – separado judicialmente
7	Qualificação anterior (grau)	contínuo	Nota da qualificação anterior (entre 0 e 200)
4	Curso	Categórica	33 - Tecnologias de Produção de Biocombustíveis. 171 - Design de Animação e Multimédia. 8014 - Serviço Social. 9003 - Agronomia. 9070 - Design de Comunicação. 9085 - Enfermagem. 9119 - Engenharia Informática. 9147 - Gestão. 9238 - Serviço Social. 9254 - Turismo. 9670 - Gestão de Publicidade e Marketing. 9773 - Jornalismo. 9991 - Gestão (presencial noturno)

Fonte: Dados originais da pesquisa.

Ademais, a referida base contém 4.424 instâncias com 36 atributos e 03 classes distintas, a saber: “Dropout”, “Graduate” e “Enrolled”. A classe “Dropout” representa os alunos que desistiram dos seus respectivos cursos dentro do intervalo do primeiro e segundo semestre. Já as classes “Graduate” e “Enrolled”, representam os alunos que conseguiram finalizar o curso e os que ainda se encontram matriculados, respectivamente.

Existe, entretanto, um forte desequilíbrio em relação a uma das classes da BD, ou seja, a classe majoritária “Graduate” representa 49,9% dos dados, enquanto que as classes “Dropout” e “Enrolled” representam 32,1% e 17,9%, respectivamente. Em virtude disso, técnicas de balanceamento das classes foram aplicadas. Por fim, as respectivas variáveis categóricas “Dropout”, “Graduate” e “Enrolled” foram representadas por valores inteiros

antes de passar pelos algoritmos de classificação, evitando-se problemas de representação, conforme apresentado na Tabela 3.

Tabela 3. Representação das classes na base de dados por valores inteiros.

Classe	"Graduate"	"Dropout"	"Enrolled"
Nova representação na base de dados	1	-1	0

Fonte: Dados originais da pesquisa.

Antes da implementação dos algoritmos de ML, foi realizada uma análise exploratória dos dados. Dos 36 atributos apresentados na BD, dois deles (26 - Unidades curriculares 1° semestre (grau) e 32 - Unidades curriculares 2° semestre (grau)) apresentavam valores muito divergentes em escala, quando comparado com os demais. Diante disso, foi necessário realizar o processo de padronização. A técnica "z-score", é um método de padronização que transforma os dados de um conjunto de dados em uma escala com média zero e desvio padrão igual a um. Essa técnica calcula a diferença entre cada atributo e a média dos dados, dividindo-a pelo desvio padrão, o que resulta em uma distribuição com média zero e variância unitária.

A padronização é uma técnica amplamente utilizada em algoritmos de ML, pois evita a predominância de atributos com maiores escalas, permitindo que todas as características tenham um impacto igual durante o treinamento. Ademais, a padronização facilita a convergência dos algoritmos, uma vez que reduz as disparidades entre os valores das diferentes variáveis e torna o espaço de atributos mais equilibrado.

Algoritmo para desbalanceamento de classes

A qualidade dos dados é uma das principais preocupações em ML. Diversos aspectos podem influenciar no desempenho de um sistema de aprendizado devido à quantidade de dados. Em bases de dados reais, dois desses aspectos estão relacionados com (i) a presença de valores desconhecidos, os quais são tratados de uma forma bastante simplista por diversos algoritmos de ML e (ii) a diferença entre o número de instâncias, ou registros de um "dataset", que pertencem a diferentes classes, uma vez que quando a diferença é expressiva, sistemas de aprendizado podem ter dificuldades em aprender o conceito relacionado com a classe minoritária.

O conjunto de dados utilizado neste trabalho possui um desbalanceamento considerável em duas classes, conforme apresentado na Tabela 4.

Tabela 4. Quantitativo de atributos em cada classe da base de dados.

Classe	"Graduate"	"Dropout"	"Enrolled"
Número de atributos	2209	1421	794
	49,9%	32,1%	17,9%

Fonte: Dados originais da pesquisa.

Duas técnicas de amostragem comumente utilizadas são o (i) "oversampling" que consiste em realizar o aumento das classes minoritárias de forma aleatória, reduzindo o desbalanceamento e (ii) "undersampling" que realiza a diminuição da classe majoritária, igualando o número de amostras das classes. Em ambos os casos existe a possibilidade de perda de dados importantes para o processo de predição dos algoritmos de ML (Machado e Emerson, 2007).

No presente trabalho, adotamos a técnica SMOTE para realizar o desbalanceamento das classes. Esse método encontra exemplos de vizinhos da classe minoritária no espaço de atributos, sintetizando um novo exemplo no espaço entre os seus vizinhos (Chawla et al., 2002), conforme ilustrado na Figura 1.

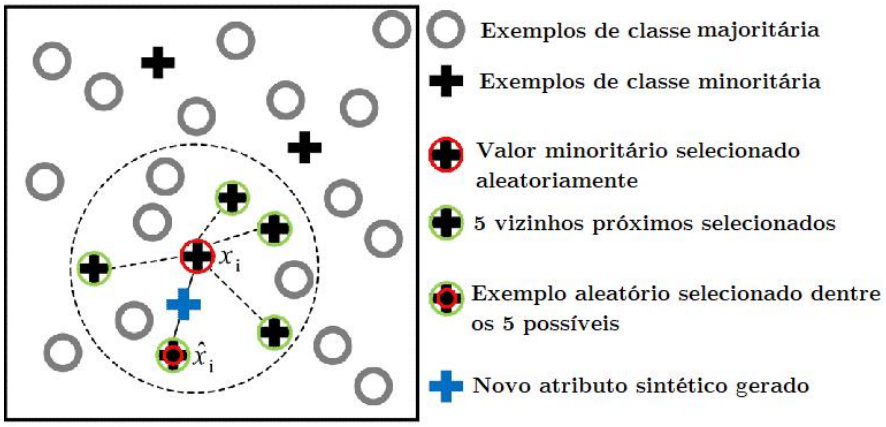


Figura 1. Geração de novos atributos sintéticos
Fonte: Elaborado pelo autor.

O cerne da técnica SMOTE envolve uma séria de equações matemáticas que descrevem como as instâncias sintéticas são geradas. Primeiramente é fundamental calcular a distância entre as instâncias no espaço n-dimensional. Neste trabalho, optou-se por utilizar a distância euclidiana, conforme apresentado na eq. (1).

$$d(A, B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \ldots + (A_n - B_n)^2} \quad (1)$$

A e B representam as duas instâncias no conjunto de dados, cada uma com n características. Em seguida, para selecionar os vizinhos mais próximos de uma instância da classe minoritária, o SMOTE utiliza o valor de k, que é um parâmetro ajustável. Neste trabalho, optou-se por utilizar k = 5. A eq. (2) representa a criação da nova instância sintética, que são interpolações ponderadas entre a instância de interesse e seus vizinhos mais próximos.

$$S = X + \alpha(X_{nn} - X) \quad (2)$$

S representa a nova instância sintética, X é a instância de interesse, X_{nn} é um dos vizinhos mais próximos selecionados e α é um valor aleatório no intervalo de 0 a 1 que controla o grau de interpolação (Chawla et al., 2002).

Algoritmos de aprendizado de máquina

O ML tem uma importância fundamental na realização de classificação e predição, tendo em vista que as máquinas usualmente realizam um processo de aprendizagem mais eficaz através dos dados apresentados durante seu processo de treinamento, teste e validação. Isso torna possível a automação de tarefas complexas e a tomada de decisões com base em dados em tempo real, o que pode levar a uma maior eficiência e produtividade em várias áreas, como saúde, transporte, dentre outras.

Neste trabalho são propostos três algoritmos baseados em ML para realizar a predição de sucesso/insucesso acadêmico de alunos, baseados nos atributos da BD disponibilizada por Valoriza (2020). Os modelos RNA-MLP, RF e DT foram propostos para verificar a eficácia de predição dos mesmos, levando-se em conta o baixo custo computacional e a facilidade de implementação. Os hiperparâmetros de cada algoritmo foram definidos através de uma busca otimizada e empírica durante a implementação dos mesmos. Todas as análises foram executadas utilizando o software “Matrix Laboratory [MATLAB]” em sua versão R2022a.

Redes Neurais Perceptron Multicamadas

Uma “Multilayer Perceptron [MLP]” é um tipo de modelo de rede neural que consiste em várias camadas ou “layers” de neurônios artificiais onde cada camada está conectada à

camada seguinte. A camada de entrada recebe os atributos, que são então processados pelas camadas ocultas, e a camada de saída fornece as previsões finais do modelo (Haykin S, 2009). A Figura 2 ilustra um modelo de RNA do tipo MLP com duas camadas ocultas, onde X_1, X_2 e X_n representam as entradas da rede.

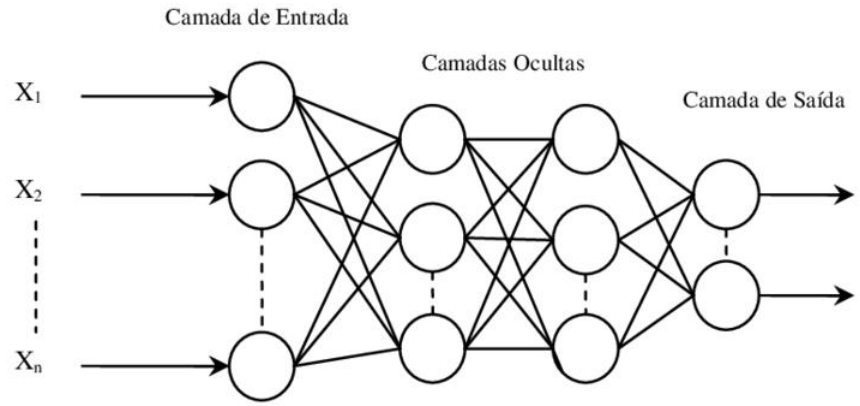


Figura 2. Modelo de uma RNA-MLP
Fonte: Adaptado de (Amorim et al., 2008).

Cada neurônio da MLP realiza uma transformação linear dos dados de entrada, seguida de uma função de ativação não linear, conforme apresentado na eq. (3) (Haykin S, 2009).

$$y = f\left(\sum_{i=1}^{m} (X_i \times W_i + b\right) \quad (3)$$

Onde X_i e W_i representam as entradas e os pesos sinápticos do i-nésimo neurônio, m representa o número de entradas do neurônio, b representa uma variável de deslocamento denominado de “bias”, f e y representam a função de ativação do neurônio e sua saída, respectivamente.

As camadas ocultas ajudam a capturar as características mais complexas e abstratas nos dados de entrada, enquanto a camada de saída produz uma previsão do modelo com base nas características aprendidas (Romelhar et al., 1986). O treinamento da MLP envolve a atualização dos pesos da rede com base em um algoritmo de otimização, como o algoritmo de retropropagação “backpropagation”, para minimizar a diferença entre as previsões do modelo e os valores reais dos dados de treinamento (Silva et al., 2019). A MLP é usada para resolver uma variedade de problemas de aprendizado de máquina, como classificação, regressão e previsão de séries temporais.

Árvore de Decisão

O modelo DT é um algoritmo de ML que é usado para problemas de classificação e regressão. A DT é uma representação visual de um conjunto de regras de decisão hierárquica que levam a uma decisão final, iniciando com um nó raiz que representa todo o conjunto de dados. Em seguida, ele é dividido em dois ou mais nós filhos, com base em uma característica de modo que cada nó filho representa uma subpopulação mais homogênea de dados do que o nó pai. Este processo é repetido recursivamente para cada nó filho até que todos os nós terminais (ou folhas) representem subpopulações puras (com todas as instâncias pertencentes à mesma classe) ou não possam mais ser divididos de maneira significativa. A árvore de decisão é uma ferramenta poderosa para entender como as decisões são tomadas em um determinado domínio e pode ser facilmente interpretada. Além disso, é possível utilizar técnicas para otimizar o tamanho da árvore e evitar o "overfitting", como a poda. Mais detalhes a respeito do modelo podem ser obtidos em (Han et al., 2011).

Floresta Aleatória

O RF é uma técnica de ML que combina várias árvores de decisão para realizar tarefas de classificação e regressão. Essa abordagem baseia-se na criação de um conjunto diversificado de árvores de decisão independentes, onde cada árvore é treinada em uma amostra aleatória dos dados de treinamento, e a previsão final é obtida por meio da média ou votação das previsões individuais das árvores. Em cada árvore da floresta, ocorre uma seleção aleatória de subconjuntos de características, o que aumenta a diversidade do conjunto de árvores. Isso contribui para reduzir o sobreajuste e aumentar a generalização do modelo. Além disso, durante a construção de cada árvore, a divisão dos nós é realizada considerando apenas um subconjunto aleatório de características, o que contribui para a independência entre as árvores.

Ademais, o RF possui vantagens como boa capacidade de generalização, resistência ao sobreajuste, tratamento eficiente de dados de alta dimensionalidade e robustez em relação a dados ausentes. É amplamente utilizado em diversos domínios, como classificação de imagens, detecção de fraudes, bioinformática e muitos outros. Mais detalhes a respeito do modelo podem ser obtidos em (Kursa e Rudnicki, 2010).

Técnicas de validação dos modelos

A utilização de técnicas de validação cruzada em algoritmos de ML é fundamental para uma avaliação robusta e confiável dos modelos. Essas técnicas permitem estimar a capacidade de generalização do modelo para dados não vistos, evitando problemas de sobreajuste. Além disso, a validação cruzada ajuda a evitar a dependência de uma única divisão de treinamento/teste, proporcionando uma avaliação mais estável do desempenho do modelo. Dessa forma, as técnicas de validação cruzada fornecem uma medida objetiva da performance do modelo, auxiliando na seleção de parâmetros, escolha do melhor modelo e na comparação de diferentes abordagens de aprendizado de máquina**.**

A técnica de validação cruzada “k-fold” é amplamente utilizada em aprendizado de máquina. Nesse método, o conjunto de dados é dividido em k partes “folds” de tamanhos aproximadamente iguais. O modelo é treinado k vezes, onde em cada iteração, um dos ‘folds” é utilizado como conjunto de teste e os k-1 “folds” restantes são usados como conjunto de treinamento. Ao final, as métricas de desempenho obtidas em cada iteração são combinadas para obter uma estimativa geral do desempenho do modelo (Géron, 2019).

Já a validação cruzada estratificada “Stratified k-fold” é uma variação da validação cruzada “k-fold” que leva em consideração a distribuição das classes durante a divisão dos “folds”. Essa técnica é especialmente útil quando o conjunto de dados apresenta um desbalanceamento significativo entre as classes. A validação cruzada estratificada mantém a proporção das classes em cada “fold”, garantindo que todos os “folds” tenham uma representação adequada de cada classe. Isso é importante para evitar viés na avaliação e garantir que o desempenho do modelo seja avaliado de forma justa em relação a todas as classes.

Neste trabalho foram realizados dois experimentos de classificação para cada algoritmo proposto. No primeiro experimento a BD foi dividida em 80% para treinamento 2 20% para teste, utilizando a técnica SMOTE para realizar o desbalanceamento das classes, fazendo com que os atributos pudessem ser utilizados diretamente nos respectivos algoritmos. Já no segundo experimento a avaliação foi feita através da validação cruzada estratificada com o desbalanceamento das classes.

Métricas de avaliação

Como vários algoritmos de classificação podem ser desenvolvidos, é importante que exista algum mecanismo de avaliação dos mesmos. Para que essa avaliação seja eficaz e fiel às características do algoritmo de classificação, é preciso cuidado na escolha das

amostras que serão utilizadas no treinamento e teste do classificador (Costa, 2008). Uma abordagem bastante popular para avaliar a performance de um modelo de classificação é baseada na Matriz de Confusão [MC].

A MC, também chamada de matriz de contingências, tem por finalidade, relacionar os resultados corretos de uma classificação com os resultados previstos pelo modelo, facilitando a visualização do número de classificações corretas e do número de classificações preditas para cada classe de um determinado conjunto de testes, segundo o classificador em análise. Torna-se uma ferramenta útil para analisar a qualidade do classificador no reconhecimento de exemplos de diferentes categorias (Castro Silva 2016).

Ainda segundo Costa 2008, para um problema de classificação binária, ou seja, quando se possui apenas duas classes, a MC pode ser utilizada conforme apresentado na Tabela 5 a seguir, onde em suas linhas são representados os objetos das classes reais e em suas colunas os objetos das classes previstas (Facelli et al., 2011).

Tabela 5. Matriz de confusão para classificação binária.

		Classe Observada	
		Positivo	Negativo
Classe Prevista	Positivo	Verdadeiro Positivo [VP]	Falso Positivo [FP]
	Negativo	Falso Negativo [FN]	Verdadeiro Negativo [VN]

Fonte: Dados originais da pesquisa.

Ao acertar a previsão da classe alvo (referida como positiva), ocorre um caso de verdadeiro positivo VP. Entretanto, ao realizar uma previsão incorreta para essa classe, ocorre um falso positivo FP. Por outro lado, quando se prevê a classe alternativa (denominada como negativa) e a classe alvo é a correta, isso é conhecido como falso negativo FN. Quando se prevê a classe alternativa corretamente, ocorre um verdadeiro negativo VN.

Para o problema multiclasse em análise com as classes “Graduate”, “Dropout” e “Enrolled” a MC pode ser representada conforme os dados apresentados na Tabela 6. Nela VP_G representa o número de VP para a classe “Graduate”. FP_G/D representa o número de FP onde a amostra pertence à classe “Graduate”, mas foi classificada como “Dropout”. FP_G/E representa o número de FP onde a amostra pertence à classe “Graduate”, mas foi classificada como “Enrolled”. VP_D representa o número de VP para a classe “Dropout”. FP_D/G representa o número de FP onde a amostra pertence à classe “Dropout”, mas foi classificada como “Graduate”. FP_D/E representa o número de FP onde a amostra pertence à

classe “Dropout”, mas foi classificada como “Enrolled”. VP_E representa o número de VP para a classe “Enrolled”. FP_E/G representa o número de FP onde a amostra pertence à classe “Enrolled”, mas foi classificada como “Graduate” e FP_E/D representa o número de FP onde a amostra pertence a classe “Enrolled”, mas foi classificada como “Dropout”.

Tabela 6. Matriz de confusão para classificação multiclasse.

		Classe Observada		
		“Graduate”	“Dropout”	“Enrolled”
Classe Prevista	“Graduate”	VP_G	FP_G/D	FP_G/E
	“Dropout”	FP_D/G	VP_D	FP_D/E
	“Enrolled”	FP_E/G	FP_E/D	VP_E

Fonte: Dados originais da pesquisa.

Para verificar o desempenho de predição dos algoritmos propostos, a acurácia [acc] foi utilizada como métrica principal de avaliação dos modelos. A acurácia informa o quanto o modelo gerado acertou em suas previsões através da soma dos verdadeiros positivos e verdadeiros negativos de cada classe, dividida pela soma total das previsões. A soma total das previsões [S_T] pode ser calculada através da eq. (4).

$$S_T = VP_G + \frac{FP_G}{D} + \frac{FP_G}{E} + VP_D + \frac{FP_D}{G} + \frac{FP_D}{E} + VP_E + \frac{FP_E}{G} + \frac{FP_E}{D} \quad (4)$$

Logo, utilizando a eq. (4), podemos calcular a acurácia para o problema multiclasse conforme apresentado na eq. (5).

$$acc = \frac{VP_G + VP_D + VP_E}{S_T} \quad (5)$$

Ademais, a métrica “F-Score [F1]”, foi utilizada como método de avaliação auxiliar dos modelos. A métrica F1 eq. (8) é dada pela média harmônica entre a precisão [P] eq. (6) e a Sensibilidade [Ss] eq. (7). As métricas Ss, P e F1 são representadas a seguir para a classe “Graduate”.

$$\mathrm{Ss_G} = \frac{\mathrm{VP_G}}{\mathrm{VP_G} + \frac{\mathrm{FP_D}}{\mathrm{G}} + \frac{\mathrm{FP_E}}{\mathrm{G}}} \quad (6)$$

$$\mathrm{P_G} = \frac{\mathrm{VP_G}}{\mathrm{VP_G} + \frac{\mathrm{FP_G}}{\mathrm{D}} + \frac{\mathrm{FP_G}}{\mathrm{E}}} \quad (7)$$

$$\mathrm{F1_G} = \frac{2 \times \mathrm{P_G} \times \mathrm{Ss_G}}{\mathrm{P_G} + \mathrm{Ss_G}} \quad (8)$$

Uma acurácia média alta indica que o modelo é capaz de classificar corretamente a maioria das amostras de teste. A precisão média alta indica que as previsões positivas do modelo são confiáveis. A sensibilidade média alta indica que o modelo é capaz de classificar corretamente a maioria das amostras positivas. Já um F1 alto indica um bom equilíbrio entre a precisão e a sensibilidade.

Resultados e Discussão

Nesta seção são apresentados os resultados obtidos com a aplicação dos modelos de ML propostos. Inicialmente, realizou-se uma análise exploratória dos dados, tendo em vista a grande quantidade de atributos da BD. A análise exploratória teve como objetivo verificar como os dados se distribuem, além de analisar a presença de “outliers”. Ademais, antes da implementação dos modelos, verificou-se o comportamento dos dados após a padronização com a técnica “z-score”. Por fim, os algoritmos de ML foram implementados, testados e validados para os dois experimentos de classificação propostos utilizando a validação cruzada estratificada juntamente com o conjunto de dados desbalanceados e a técnica SMOTE para realizar o balanceamento das classes minoritárias.

Padronização dos Dados

A padronização é uma etapa essencial no pré-processamento de dados antes da aplicação dos algoritmos de ML. Essa técnica tem como objetivo garantir que todas as variáveis estejam na mesma escala, evitando que algumas variáveis com valores mais altos dominem as outras durante o treinamento do modelo.

Os valores dos atributos originais da BD variam entre si, em alguns casos apresentando “outliers”. Diante disso, realizou-se uma análise da distribuição dos valores de

cada atributo através de seus respectivos histogramas, conforme ilustrado na Figura 3 a seguir, onde o eixo x representa o valor do atributo e o eixo y a sua respectiva frequência.

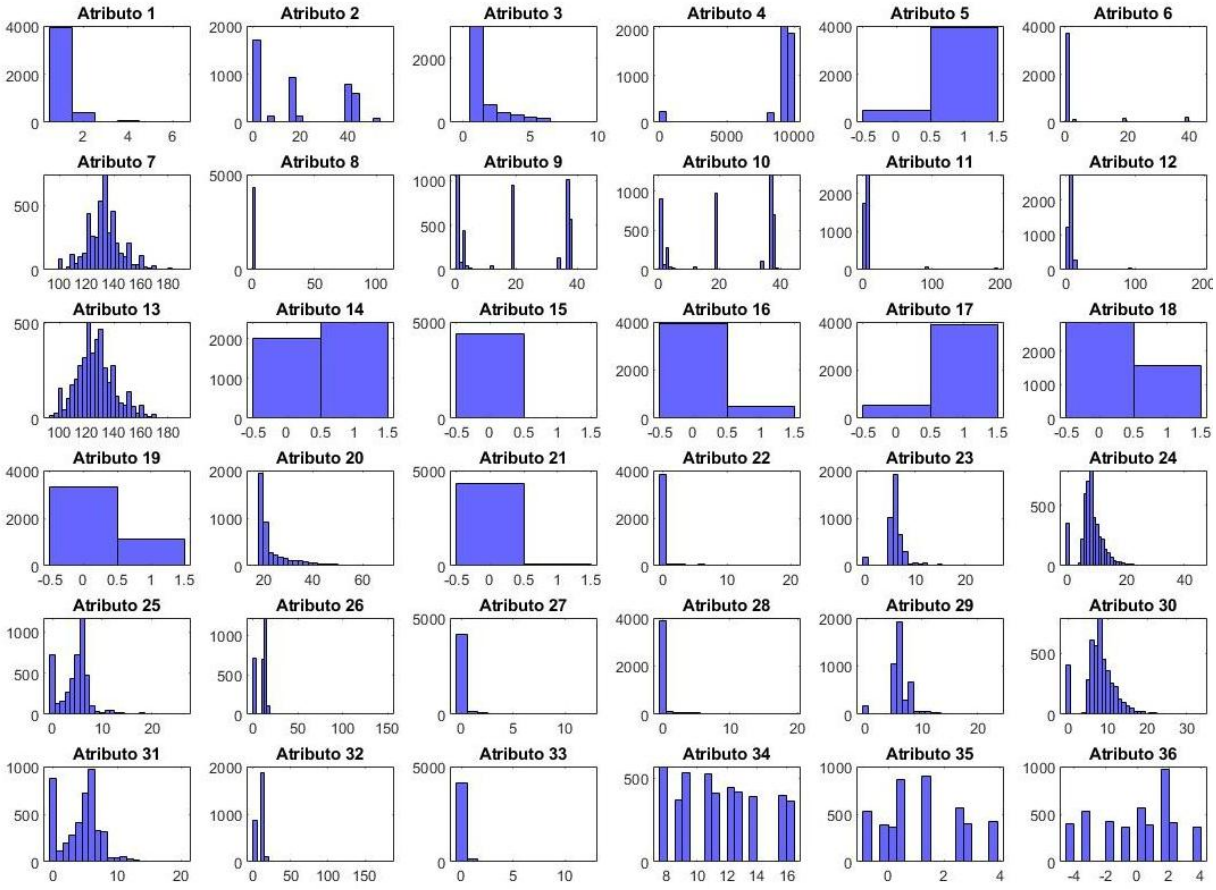


Figura 3. Histogramas das frequências dos valores de cada atributo
Fonte: Resultados originais da pesquisa.

A análise dos histogramas pode fornecer “insights” sobre a distribuição dos atributos e ajudar a identificar características relevantes dos dados. Ademais, ao observar os histogramas, é possível verificar se os atributos possuem distribuições simétricas, assimétricas, se estão concentrados em torno de um valor específico ou se apresentam uma dispersão.

Após a padronização dos dados com a técnica “z-score”, os valores de cada atributo da BD se manteve em uma escala mais próxima, conforme ilustrado na Figura 4.

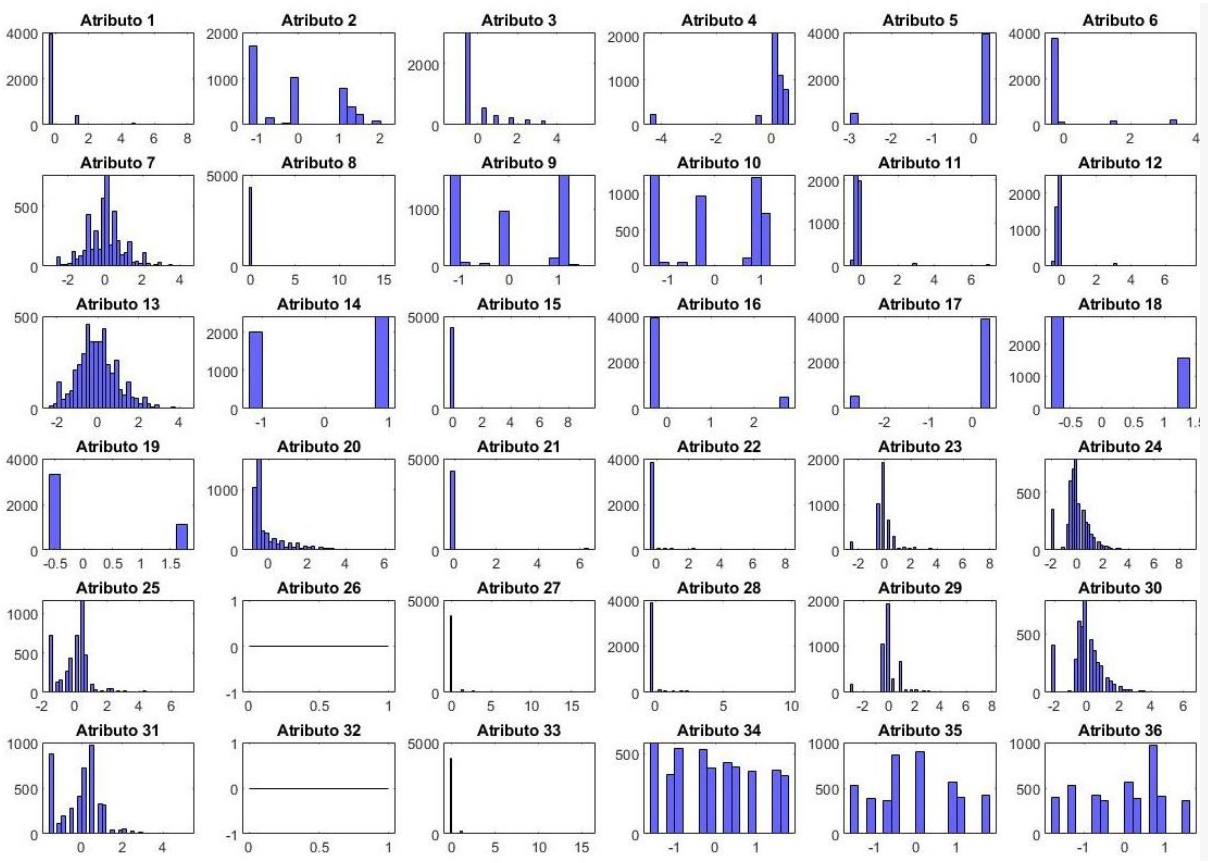


Figura 4. Histogramas das frequências dos valores de cada atributo padronizado.
Fonte: Resultados originais da pesquisa.

Verificamos também que os Atributos (26 - Unidades curriculares 1° semestre (grau) e 32 - Unidades curriculares 2° semestre (grau)), apresentam uma frequência bem próxima de zero. Diante disso, optou-se por retirá-los da base. Após a análise e padronização dos dados, a BD resultante apresenta 34 atributos.

Balanceamento de classes com a técnica SMOTE

A BD utilizada neste trabalho possui um desbalanceamento na classe 0, quando comparada com as classes -1 e 1, conforme ilustrado na Figura 5. Ademais, existe um desbalanceamento entre a classe -1 e 1, apresentando 14121 e 2209 instâncias, respectivamente.

Figura 5. Histograma do número de instâncias das classes
Fonte: Resultados originais da pesquisa.

A técnica SMOTE cria instâncias sintéticas da classe minoritária, aumentando sua representação, realizando o processo de balanceamento das classes. Para implementar a técnica SMOTE, primeiro se faz necessário definir o número de vizinhos a serem considerados na comparação das instâncias para a criação das novas, neste caso, o valor de k foi escolhido empiricamente igual a 5. A Figura 6 apresenta a quantidade de instâncias de cada classe após o balanceamento com a técnica SMOTE.

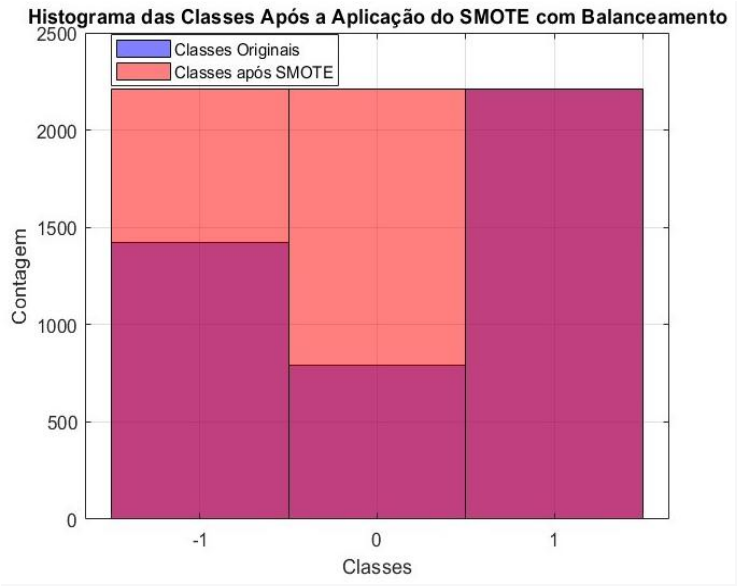


Figura 6. Histograma do número de instâncias das classes após balanceamento com SMOTE
Fonte: Resultados originais da pesquisa.

Modelagem preditiva com validação cruzada estratificada

Inicialmente, para a avaliação das métricas de desempenho dos modelos propostos, foram realizados os treinamentos com a BD desbalanceada, utilizando a validação cruzada estratificada.

Cada algoritmo realizou a validação cruzada através de 5 “folds”. As métricas de acurácia e F1 foram obtidas para cada “fold” e, em seguida foram calculadas as suas respectivas médias, fornecendo uma visão geral do desempenho de cada modelo. Assim sendo, após as análises dos algoritmos por meio da validação cruzada estratificada, foram extraídos os seguintes resultados, conforme apresentado na Tabela 7.

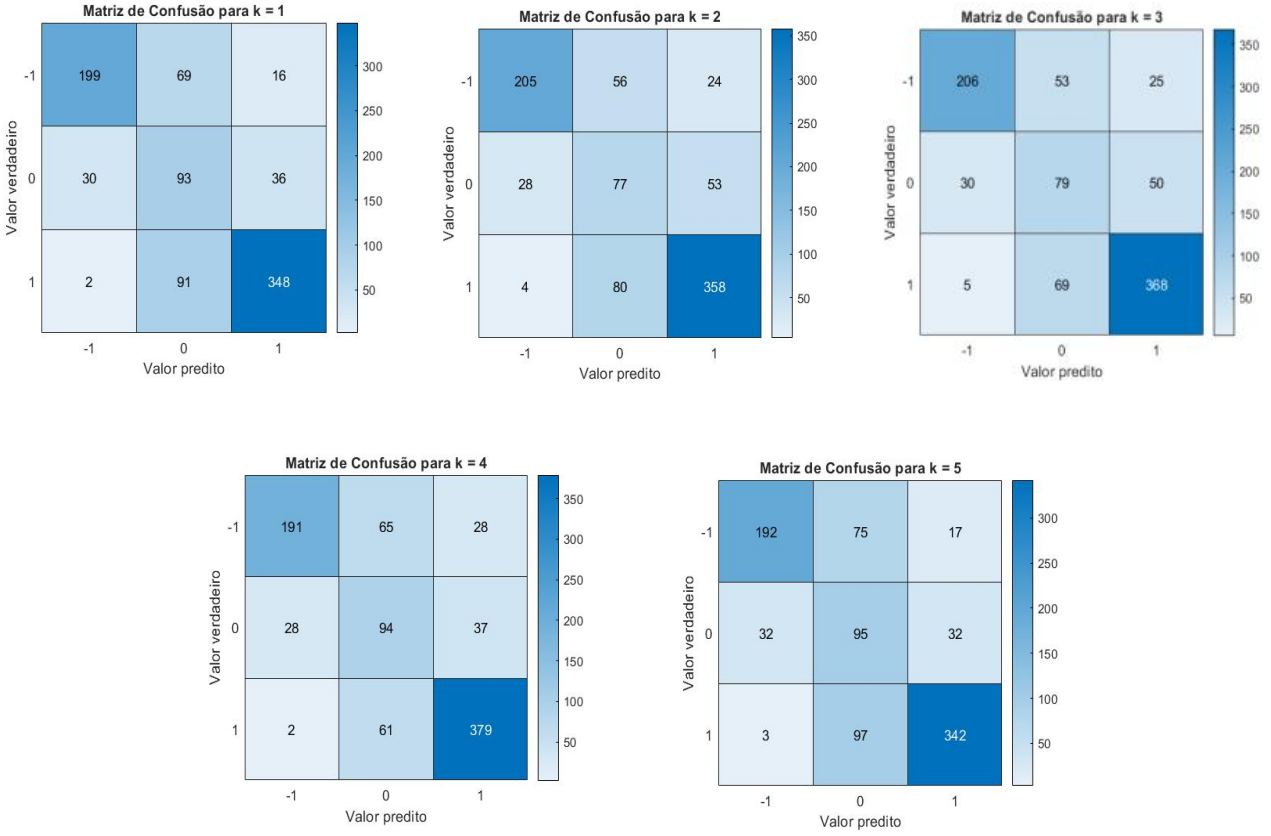
Tabela 7. Média das métricas dos modelos sem balanceamento com validação cruzada estratificada.

Modelo	Acurácia	F1
RNA-MLP	73,15%	85,87%
Random Forest	78,03%	67,30%
Decision Tree	70,30%	80,12%

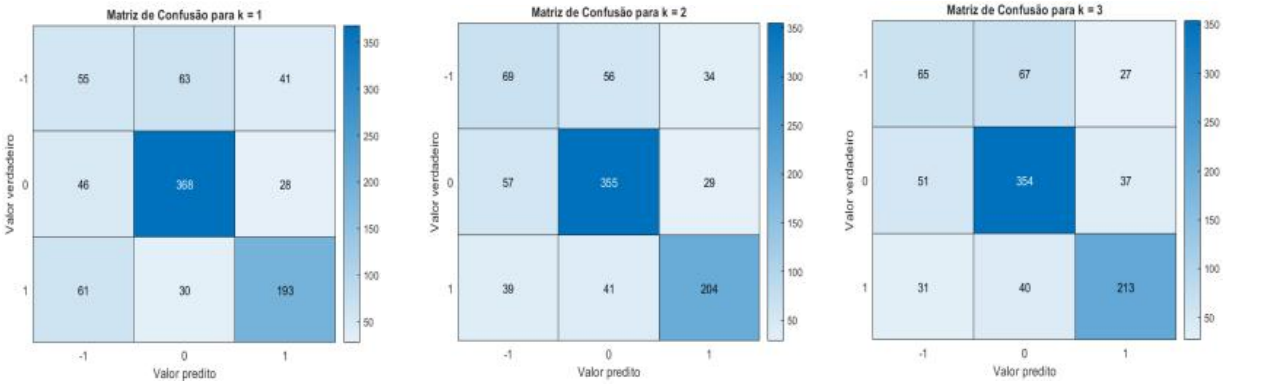
Fonte: Resultados originais de pesquisa.

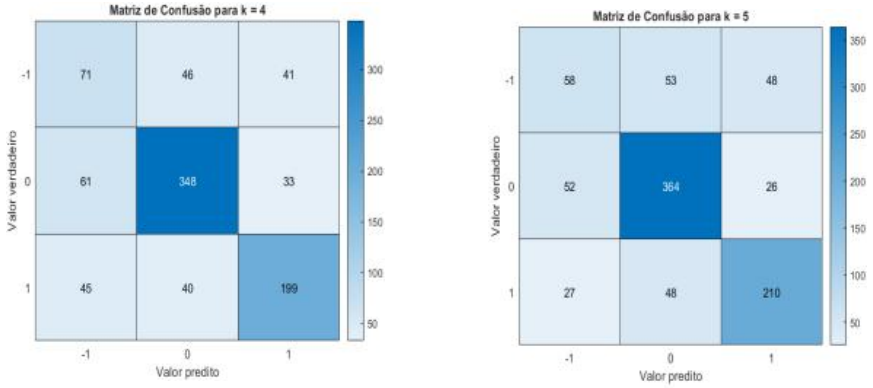
Com base nos resultados apresentados para os três modelos, pode-se fazer as seguintes observações e análises: O RNA-MLP obteve uma acurácia média de 73,15%, o que significa que o modelo classificou corretamente 73,15% das amostras de teste; comparando com o modelo RF, a RNA-MLP apresentou uma acurácia inferior, estando superior ao modelo DT que obteve uma acurácia média de 70,30%. O RF obteve a maior acurácia dentre os três modelos, indicando que conseguiu classificar corretamente a maior proporção de amostras de teste. Por fim, a RNA-MLP obteve o maior F1 (85,87%), seguidos dos modelos DT e RF com 80,12% e 67,30%, respectivamente. Isso indica que a RNA-MLP conseguiu alcançar um equilíbrio entre a precisão e a sensibilidade melhor do que os outros dois modelos.

A Figura 7(a), 7(b) e 7(c) apresenta as matrizes de confusão dos modelos RNA-MLP, DT e RF para cada iteração da validação cruzada estratificada.

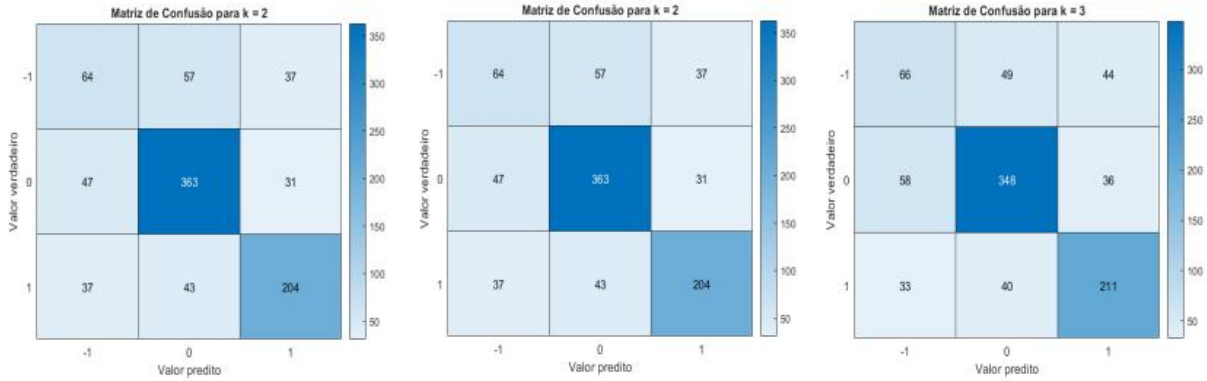

Matriz de Confusão para k = 1
Valor verdadeiro
Valor predito
-1
0
1
199
69
16
30
93
36
2
91
348
Matriz de Confusão para k = 2
205
56
24
28
77
53
4
80
358
Matriz de Confusão para k = 3
206
53
25
30
79
50
5
69
368
Matriz de Confusão para k = 4
191
65
28
28
94
37
2
61
379
Matriz de Confusão para k = 5
192
75
17
32
95
32
3
97
342
50
100
150
200
250
300
350

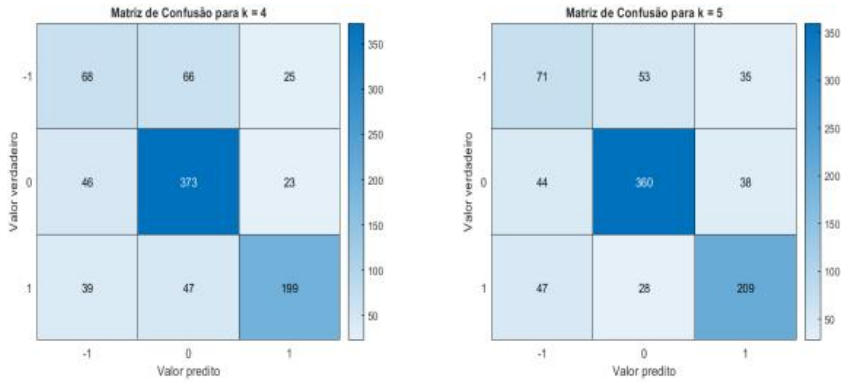
(a)


Matriz de Confusão para k = 1
Valor verdadeiro
Valor predito
-1
0
1
55
63
41
46
368
28
61
30
193
Matriz de Confusão para k = 2
69
56
34
57
365
29
39
41
204
Matriz de Confusão para k = 3
65
67
27
51
354
37
31
40
213
50
100
150
200
250
300
350



(b)





(c)

Figura 7. Resultado das Matrizes de Confusão para a validação cruzada estratificada
Fonte: Resultados originais da pesquisa.

Verifica-se na Figura 7(a) que o modelo RNA-MLP realizou uma predição assertiva para as classes -1 e 1, realizando um baixo número de classificações corretas para a classe 0. Ademais, a classe -1 (“Droupout”) representava apenas 32,1% do total de amostras da BD. Já os modelos DT e RF realizaram uma predição assertiva para as classes 0 e 1. Embora a classe 0 (“Enrolled”) estivesse representada apenas por 17,9% das amostras, os

algoritmos conseguiram realizar uma classificação assertiva por meio da validação cruzada estratificada.

Modelagem preditiva com balanceamento de classes através da técnica SMOTE

Para que fosse possível realizar o balanceamento do conjunto de dados, primeiramente, os atributos e as classes foram separados da BD. Em seguida, verificou-se o desbalanceamento das classes, contando o número de amostras em cada classe da base.

O SMOTE foi aplicado para cada classe minoritária presente nos dados. Diante disso, foi necessário fazer uma varredura que percorresse cada classe, verificando se o número de amostras era menor que o número máximo de amostras desejado. Se fosse o caso, amostras sintéticas seriam geradas para equilibrar a classe.

A Tabela 8 apresenta os resultados dos algoritmos implementados anteriormente no conjunto de teste utilizando a base balanceada através da técnica SMOTE.

Tabela 8. Métricas dos modelos com balanceamento das classes através da técnica SMOTE.

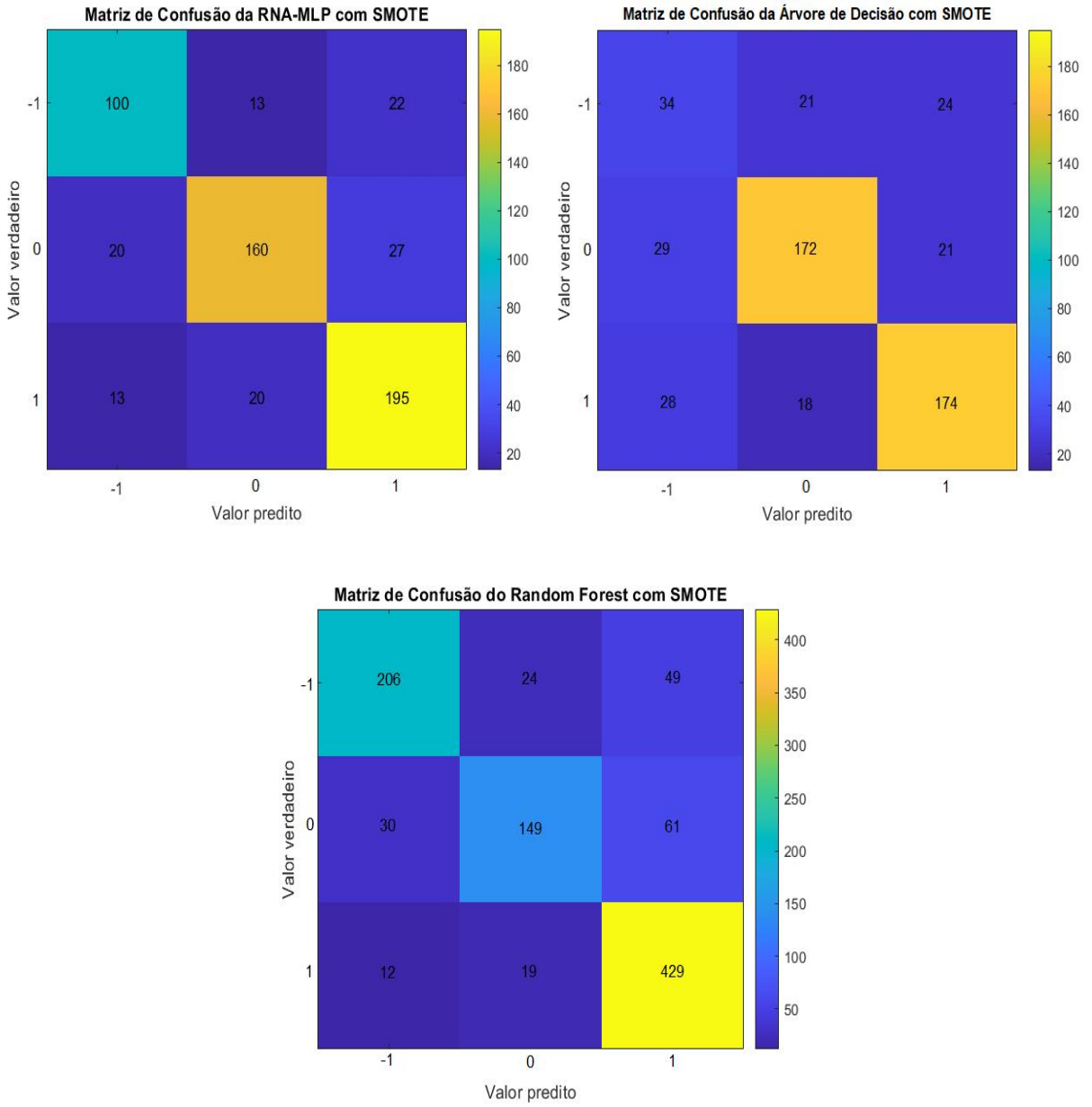
Modelo	Acurácia	F1
RNA-MLP	80,42%	87,60%
Random Forest	80,68%	87,78%
Decision Tree	74,04%	83,26%

Fonte: Resultados originais de pesquisa.

Verificando inicialmente a acurácia, o modelo RF apresentou o melhor resultado, com 80,68%, seguido da RNA-MLP com 80,42% e DT com 74,04%. Em relação a métrica F1 o modelo RF mais uma vez se destacou, com 87,78%, seguido pelo RNA-MLP com 87,60% e DT com 83,26%.

Com base nos dados apresentados, pode-se observar que cada modelo teve seu destaque em uma métrica específica. O modelo RF se destacou pela maior acurácia e F1, indicando um bom desempenho global. Por outro lado. O modelo DT apresenta resultados intermediários nas métricas avaliadas.

A Figura 8(a), 8(b) e 8(c) apresenta as matrizes de confusão dos modelos RNA-MLP, DT e RF para cada iteração da validação cruzada estratificada.



(a) (b) (c)

Figura 8. Resultado das Matrizes de Confusão para a validação utilizando os dados balanceados com a técnica SMOTE
Fonte: Resultados originais da pesquisa.

Por meio das matrizes de confusão apresentadas na Figura 8 acima, verifica-se que os modelos RNA-MLP e RF obtiveram uma predição assertiva nas três classes, enquanto que a AD, não realizou um bom processo de predição para a classe -1.

Considerações Finais

Neste trabalho, foram investigados a utilização de algoritmos de aprendizado de máquina para predizer o desempenho acadêmico de alunos de cursos superiores, visando apresentar “insights” para as instituições de ensino para que medidas preventivas sejam tomadas antes da desistência dos alunos.

Para isso, três modelos de aprendizado de máquina foram propostos e dois experimentos foram realizados em uma BD desbalanceada. No primeiro experimento os modelos foram treinados através da validação cruzada estratificada, dividindo os dados desbalanceados em 5 “folds”. Já no segundo experimento, utilizamos a técnica SMOTE para gerar instâncias sintéticas, balanceando a base de dados e dividindo os dados em 80% para treinamento e 20% para teste. Considerando os resultados obtidos pelos três modelos, verificamos que o RF demonstrou melhor desempenho global, obtendo a maior acurácia média, quando comparado com o RNA-MLP e DT. Ademais, constamos que a técnica SMOTE obteve uma melhora significativa da acurácia em ambos os modelos. Essas observações ressaltam a importância de selecionar o modelo mais apropriado com base nas métricas relevantes para a aplicação específica do problema em questão.

Agradecimento

A Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo [ESALQ/USP] pela concessão da bolsa integral de estudos. À minha orientadora Caroline Belisário Zorzal pelos conhecimentos compartilhados. Meus sinceros agradecimentos.

Referências

Martins, M.V., Tolledo, D., Machado, J., Baptista, L.M.T., Realinho, V. (2021). Early Prediction of student’s Performance in Higher Education: A Case Study. In: Rocha, Á., Adeli, H., Dzemyda, G., Moreira, F., Ramalho Correia, A.M. (eds) Trends and Applications in Information Systems and Technologies. WorldCIST 2021. Advances in Intelligent Systems and Computing, vol 1365. Springer, Cham.

Mduma, N., Kalegele, K., Machuve, D.: A survey of machine learning approaches and techniques for student dropout prediction. Data Sci. J. 18, 1–10 (2019).

Beaulac, C., Rosenthal, J.S.: Predicting university Students’ academic success and major using random forests. Res. High. Educ. 60, 1048–1064 (2019.

Hoffait, A.S., Schyns, M.: Early detection of university Students with potential difficulties. Decis. Support Syst. 101, 1–11 (2017).

Miguéis, V.L.; Freitas, A., Garcia; P.J.V; Silva, A. 2018. Early segmentation of students according to their academic performance: a predictive modelling approach. Decis. Support Syst. 115.

Thammasiri, D.; Delen, D.; Meesad, P.; Kasap, N. 2014. A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition. Expert Syst. Appl. 41.

Chawla, N.V.; Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell.

Romero, C., Ventura, S. 2010. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). Part C Ap. Rev.40.

Valoriza — Research Center for Endogenous Resource Valorization, Instituto Politécnico de Portalegre, 7300-555 Portalegre, Portugal Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Portalegre, 7300-555 Portalegre, Portugal.

Vinicius, A., Sobreio, Marcelo, S., Nagano. 2008. Uma Estimação do Valor da Commodity de Açucar Utilizando Redes Neurais Artificiais. Revista P&D em Engenharia de Produção. N° 7 (2008) p. 36-52.

Guillaume L; Fernando N.; e Christos K. Aridas. 2017. Inbalanced-learn: A Python Toolbox to Trackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research.

Naser, M. Z.; ALAVI, A. 2020. Insights into performance fitness and error metrics for machine learning. CoRR, abs/2006.00887.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986. Learning representations by back-propagating errors. Nature, 323(6088).

Kursa, M. B., e Rudnicki, W. R. 2010. Feature selection with the Boruta package. Journal of statistical software, 36(11).

Han, J., Kamber, M., & Pei, J. 2011. Data mining: concepts and techniques. Elsevier.

Haykin, S. S. Neural networks and learning machines. Third. Upper Saddle River, NJ: Pearson Educacion, 2009.