Análise de Classificadores para Predição de Evasão em um Curso Superior de Tecnologia da Informação de um Instituto Federal

Anais do Computer on the Beach

Felipe Ferreira de Sousa
Instituto Federal de Brasília - IFB, Brasil
felipe57121@estudante.ifb.edu.br

Eduardo Gabriel Ferreira Silva
Instituto Federal de Brasília - IFB, Brasil
eduardo.silva6@estudante.ifb.edu.br

Claudio Ulisse
Instituto Federal de Brasília - IFB, Brasil
claudio.ulisse@ifb.edu.br

Fernando Wagner Brito Hortêncio Filho
Instituto Federal de Brasília - IFB, Brasil
fernando.filho@ifb.edu.br

Abstract

School dropout has been a long-standing issue in Higher Education Institutions (HEIs), raising concerns and prompting mitigation efforts. This paper examines dropout in a Technology course at a Federal Institute, addressing the research question: "How can machine learning classifiers assist in predicting student dropout?". The study applies machine learning classifiers using data from a Federal Institute, including demographic (age group, income, city, special needs, ethnicity, gender) and academic variables (admission method, vacancy type, enrollment status). The objective is to develop analyses ranging from descriptive statistics to predictive modeling. The classifiers used include Support Vector Machine, Random Forest, Logistic Regression, Gradient Boosting, AdaBoost, and K-Nearest Neighbors. The goal is to support teachers, coordinators, and administrators in implementing preventive measures such as personalized mentoring, continuous monitoring, and institutional policies to improve academic infrastructure and inclusivity. The results show that Support Vector Machine, Gradient Boosting, and AdaBoost achieved the best performance, with F1-measure values between 0.6 and 0.8, demonstrating their predictive capability. This highlights the potential of machine learning in addressing student dropout in higher education.

Keywords

Evasão Discente, Ensino Superior, Aprendizagem de Máquina, Mineração de Dados.

1 Introdução

A evasão escolar é um fenômeno preocupante para as Instituições de Ensinos Superiores (IES) e nos últimos anos vem se mostrando como algo que necessita de atenção dos órgãos governamentais responsáveis pela educação, para que seja possível analisar e tomar medidas com intuito de combater esse índice que vem se tornando cada vez maior. De acordo com dados do painel estatístico do Censo da Educação Superior disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), é possível observar que a taxa de evasão dos alunos no ensino superior vem apresentando um aumento alarmante no período de 2018 a 2023, alcançando 60% no ano de 2023 [1]. A Figura 1 ilustra o aumento dessa taxa ao longo dos anos.

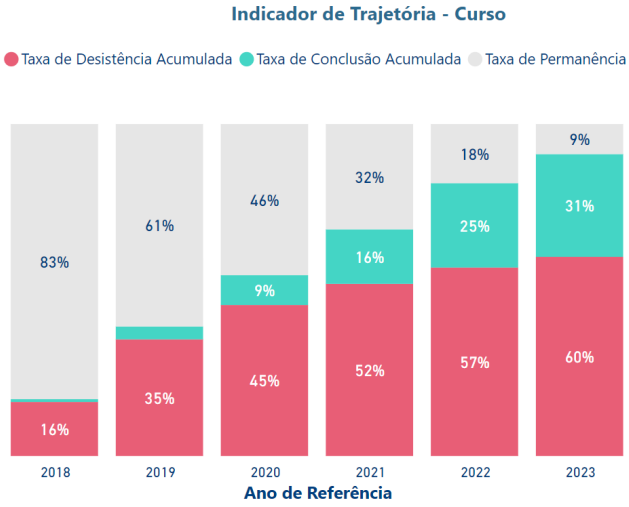

Figura 1: Painel Estatístico, via Censo da Educação Superior, (2024).

Esse fenômeno pode ocorrer desde o ensino básico até o ensino superior, gerando assim, grandes impactos negativos na qualidade de vida de um discente. Ele não só afeta a formação acadêmica do indivíduo, mas também pode interferir em oportunidades de emprego com melhores salários.

Em um levantamento realizado em [2], utilizando dados de 2009 a 2022 disponibilizados pelo o INEP foi observado que para a modalidade presencial dos ingressantes para cursos superiores, foram registrados uma taxa de 30% de alunos evadidos de instituições de redes privadas e 23,4% de redes públicas. Para a modalidade de ensino a distância (EAD), as taxas foram de 40% para instituições de redes privadas e 31,1% de redes públicas. Assim, é crucial desenvolver métodos ou estratégias para impedir que esse índice continue aumentando. Nesse contexto, a proposta de um modelo preditivo utilizando algoritmos de aprendizado de máquina que visa prever possíveis evasões de estudantes na instituição.

Ao utilizar os algoritmos de aprendizado de máquina (*Machine learning*), é possível trabalhar com uma análise preditiva de dados. Diversas empresas utilizam este conceito em prospecções e tomadas de decisões estratégicas [3]. Como exemplo, uma instituição financeira pode utilizar aprendizado de máquina para coletar predição sobre seus clientes, com essas informações, eles podem criar insights que podem ajudar a prever se um determinado cliente deixará

de pagar um empréstimo, quais clientes podem ser mais lucrativos para direcionar recursos e gastos com marketing. Outro exemplo é na área da saúde, na qual a aprendizagem de máquina pode ser utilizada para detectar e gerenciar o atendimento aos pacientes com doenças crônicas e outros acompanhamentos importantes que podem ser monitorados [4].

Na educação, a aprendizagem de máquina pode ser utilizada para analisar a relação entre evasão escolar e a autoavaliação institucional [5]. Outro exemplo é melhorar a experiência de aprendizagem do aluno, enriquecendo os instrumentos de ensino e de avaliação com características mais atrativas [6].

Diante deste cenário, pretende-se realizar uma análise dos dados disponibilizados dos alunos ingressantes desse curso de Tecnologia de um Instituto Federal. Com essa análise, através de técnicas de aprendizagem de máquina, pretende-se auxiliar servidores e gestores na prevenção da evasão discente com base nos dados e resultados alcançados.

2 Fundamentação teórica

2.1 Problema da Evasão escolar

A evasão escolar é um problema presente em todos os níveis de ensino formal. Entende-se que a evasão escolar está associada ao desligamento do discente da instituição de ensino, independentemente do motivo, exceto em casos de conclusão ou diplomação do curso [7]. De acordo com [8], a evasão em cursos superiores pode ser caracterizada pelo desligamento do aluno em decorrência de abandono, o que pode se dar por situações como a não realização da matrícula, transferência para outra instituição de ensino, mudança de curso, trancamento ou exclusão por descumprimento de alguma norma institucional. O trabalho descrito em [9] identifica dois tipos de evasão: a voluntária e a involuntária. A evasão voluntária ocorre por iniciativa do aluno, enquanto a evasão involuntária, de caráter compulsório, ocorre por intervenção da instituição de ensino superior (IES) devido a diversas razões. Independentemente da categoria, ambas são contabilizadas nas taxas de evasão.

A evasão pode acontecer por diversas causas, porém segundo [10] podemos classificá-las em externas e internas. As externas dizem respeito a questões da instituição, corpo docente, metodologias, escolha inadequada do curso e etc. Já as internas representam as necessidades mais pessoais dos sujeitos, tais como condições econômicas, problemas psicológicos e familiares. Outro ponto que pode gerar um abandono do curso é o equívoco do discente no momento de escolher o curso pretendido. Raramente um candidato a uma vaga em curso universitário tem informações completas sobre a carreira pretendida. O estudante do último ano do Ensino Médio ou de cursinho preparatório decide-se por determinado curso, mas nem sempre essa escolha é a correta [11].

No plano estratégico para permanência e êxito (PPE) do Instituto Federal do Ceará (IFCE), [12], os autores classificaram as causas de evasão em três categorias: fatores individuais, internos e externos. Os fatores individuais estão relacionados a particularidades específicas do próprio discente, como, por exemplo, a dificuldade de adaptação ao curso escolhido, falta de tempo para estudar, dificuldade em conciliar estudos e trabalho, problemas de saúde, entre outros. Os fatores internos referem-se a questões institucionais, de ordem pedagógica ou administrativa, que influenciam direta ou indiretamente o discente. Exemplos incluem sobrecarga de disciplinas, conteúdos excessivamente extensos e paralisações como greves. Por fim, os fatores externos dizem respeito a questões econômicas, sociais e do mercado de trabalho que também impactam a permanência e o êxito acadêmico. Alguns exemplos são a distância entre o campus e a residência do discente, a desvalorização da profissão docente e da de tecnólogo pela sociedade, além de dificuldades ou falta de transporte local e intermunicipal que facilite o acesso à instituição.

A evasão pode ser definida como a não conclusão do curso por qualquer motivo. Um aluno é considerado evadido quando realiza a matrícula, mas não inicia as atividades propostas, ou quando inicia o curso e desiste antes de cumprir todos os requisitos para obtenção do diploma [13]. Na educação superior, a evasão acarreta prejuízos econômicos, sociais e culturais para as instituições, que sofrem perdas financeiras e reduzem sua contribuição para o desenvolvimento da sociedade [14]. Além disso, a evasão não é vista apenas como uma dificuldade individual do aluno, mas como um reflexo de falhas estruturais da sociedade, incluindo deficiências nas instituições educacionais e na atuação dos educadores. Isso impacta diretamente os resultados escolares e aumenta os custos financeiros e sociais para famílias e o Estado [15].

2.2 Machine Learning para classificação

Aprendizado de máquina é um ramo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos capazes de aprender a partir de dados e realizar tarefas específicas sem serem explicitamente programados para isso [16]. De acordo com [17], "*Machine Learning* é o campo de estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados". Essa definição destaca a essência do aprendizado de máquina, que é a automação do aprendizado a partir de experiências passadas.

Com a crescente quantidade de dados disponíveis, analisar, extrair informações e gerar predições tem se tornado um desafio. Nesse contexto, [18] explicita que modelos de aprendizagem de máquina podem facilitar tais tarefas.

Um algoritmo de *Machine Learning* (ML) pode ser dividido em forma supervisionada ou não supervisionada. O aprendizado supervisionado é baseado no treinamento de uma amostra de dados com a classificação correta já atribuída (também chamados de rótulos), enquanto que o não supervisionado se refere à capacidade de aprender e organizar informações sem a atribuição da classificação correta [19].

O aprendizado supervisionado é uma abordagem de aprendizado de máquina onde um modelo é treinado utilizando um conjunto de dados rotulado. Esse conjunto é geralmente dividido em duas partes: um conjunto de treino e um conjunto de teste. O conjunto de treino é utilizado para ensinar o modelo, ajustando seus parâmetros para que ele consiga realizar previsões precisas. Já o conjunto de teste serve para avaliar a performance do modelo, verificando sua capacidade de generalizar para novos dados.

Essa técnica pode ser aplicada na criação de modelos preditivos para identificar o risco de evasão discente. A partir de dados históricos sobre o desempenho dos alunos, como notas, frequência e engajamento, é possível treinar um modelo supervisionado para prever quais alunos têm maior probabilidade de evadir, permitindo

ações preventivas para mitigar esse risco.[20]. As próximas seções apresentam os algoritmos supervisionados selecionados para a construção desta pesquisa.

2.3 Trabalhos relacionados

Diante desse contexto, considerando a crescente taxa de utilização de algoritmos de aprendizagem de máquina apresentada, esta seção discutem os trabalhos relacionados a esse tema.

A pesquisa descrita em [5] investiga a aplicação de algoritmos de aprendizado de máquina para prever a evasão escolar. Para a análise, foram utilizados dados de um modelo semestral de autoavaliação dos cursos de graduação da Universidade Federal da Paraíba (UFPB). O estudo comparou o desempenho de Árvore de Decisão, Floresta Aleatória e Máquinas de Vetores de Suporte (SVM), utilizando dados desbalanceados, balanceados por subamostragem aleatória e balanceados por sobreamostragem SMOTE. Como resultado, o algoritmo Floresta Aleatória se destacou, especialmente com dados balanceados por SMOTE [21], apresentando uma acurácia de 87,97%, precisão de 91,72% e sensibilidade de 91,67%. O modelo indicou que 59% dos alunos ativos tinham potencial de evasão, refletindo as taxas atuais na Paraíba. Os autores sugerem explorar modelos específicos por curso e a incorporação de dados adicionais, além de desenvolver uma API e dashboards interativos para aprimorar a predição e fornecer insights em tempo real.

O trabalho desenvolvido em [22] utilizou os dados disponibilizados o curso de "Informática" da *Hellenic Open University* (HOU) para estudar se o uso de técnicas de aprendizado de máquina pode ser útil no tratamento do problema de evasão discente. Para o desenvolvimento dos estudos foram selecionados as seis técnicas de aprendizado de máquina mais comuns, nomeadamente árvores de decisão, redes neurais, algoritmo Naive Bayes, algoritmos de aprendizagem baseados em instâncias, Regressão Logística e Máquinas de Vetores de Suporte. Após a avaliação dos modelos, foi constatado que precisão chegou a 63% nas previsões iniciais baseadas apenas em dados demográficos dos alunos e ultrapassa 83% antes do meio do período acadêmico [22].

3 Metodologia

Para auxiliar no desenvolvimento deste trabalho, foi utilizada a metodologia *CRoss-Industry Standard Process for Educational Data Mining* (CRISP-EDM) [23]. Essa metodologia foi adaptada do processo CRISP-DM (acrônimo de CRoss-Industry Standard Process for Data Mining) para melhorar o contexto da mineração de dados educacionais. A Figura 2 exibe as etapas da metodologia CRISP-EDM.

O processo do CRISP-EDM é dividido em seis etapas, todas estas voltadas ao domínio educacional. Segundo [23], as etapas não precisam ser seguidas obrigatoriamente na ordem proposta e, se necessário, podem ocorrer transições entre elas ou até mesmo ser divididas em etapas menores ou aglutinadas em etapas maiores para um melhor desenvolvimento do modelo proposto. A seguir, destacam-se como as etapas foram organizadas ao longo do presente trabalho: As etapas de Entendimento do Domínio em Educação e Entendimento dos Dados Educacionais foram combinadas e abordadas na seção Coleta e Compreensão dos Dados. Em seguida, a fase de pré-processamento de dados é apresentada na seção 3.2. As etapas de

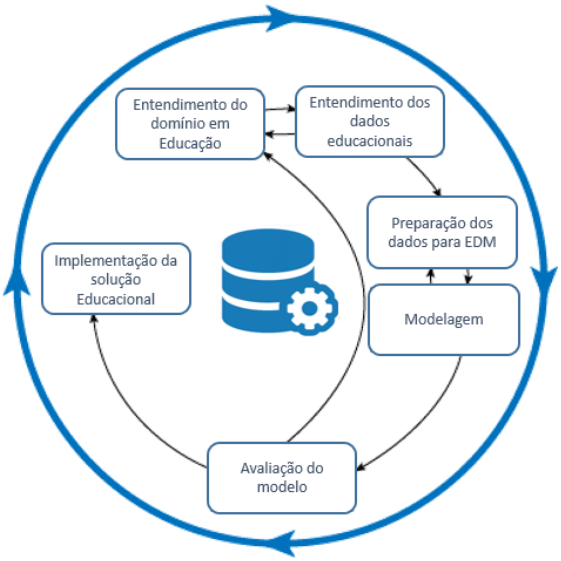

Figura 2: Metodologia CRISP-EDM. via Jorge Ramos, Rodrigo Rodrigues, João Silva, and Pamella Oliveira[23]. (2020).

geração e avaliação dos modelos foram tratadas nas seções 3.4 e 3.5 respectivamente. Por fim, a última etapa, referente à Implementação da Solução Educacional consta como trabalho futuro a ser desenvolvido posteriormente.

3.1 Coleta e compreensão dos dados

Inicialmente, foi realizado o levantamento e a análise da problemática. Como estudo de caso, foram utilizados dados fornecidos pelo setor acadêmico de uma instituição federal, abrangendo os alunos ingressantes em um curso superior de Tecnologia da Informação. A base de dados contempla todos os estudantes que ingressaram no curso desde o primeiro semestre de 2017 até 7 de novembro de 2023, data em que a extração foi realizada. No total, foram analisados dados de 879 alunos, considerando tanto aqueles que ainda estão matriculados quanto os que já passaram pelo curso nesse período.

A Tabela 1 exibe as variáveis extraídas de cada discente do curso. Considerando que fatores sociais como etnia, idade, gênero, renda per capita, cidade onde reside, tipo de vaga, dentre outros, podem influenciar na evasão, entende-se que estas são variáveis importantes tanto nas análises iniciais quanto na geração de um modelo de predição.

Já a variável a ser predita neste trabalho (situação de matrícula), possuía inicialmente os seguintes valores, conforme a Tabela 2. De acordo com [24], qualquer aluno ingressante que não conclui o curso é evadido. Logo, este conceito inclui as situações "Evadido", "Cancelado", "Transferido Externamente" e "Transferido Internamente". A situação "Transferido Externamente" se refere ao aluno que saiu do âmbito da instituição original e se matriculou em outra instituição. Já a situação "Transferido internamente" se refere ao aluno que transferiu a matrícula de um curso para um outro curso dentro da mesma instituição. Nestes casos de transferência, apesar

Tabela 1: Dados

Dado Bruto	Descrição
Nascimento	Data de nascimento
Sexo	Gênero (masculino/feminino)
Etnia	Grupo cultural
Renda per capita familiar	Renda por pessoa da família
Cidade	Local de residência
Necessidade específica	Pessoa com deficiência
Tipo de vaga	Cota ou ampla concorrência
Situação matrícula	Status da matrícula
Ano de ingresso	Ano que entrou na instituição
Semestre de ingresso	Semestre que se encontra

Tabela 2: Dados de situação matrícula

Situação de matrícula
Cancelado
Concluído
Evadido
Matriculado
Trancado
Transferido externamente
Transferido internamente

de contar como aluno ingressante no novo curso, esse mesmo aluno conta como aluno evadido no curso anterior.

3.2 Pré-processamento dos dados

Neste estágio, os dados foram transformados em dados preparados, para serem padronizados de acordo com o método, e assim facilitar a modelagem na fase seguinte da metodologia CRISP-EDM. Essa é uma etapa muito importante, pois nela é possível identificar os dados que fazem sentido para a metodologia escolhida. Ela é composta por limpeza de valores nulos, padronização de dados que podem ser correlacionados, criação de novas colunas ou variáveis que podem ser essenciais para análise. Para auxiliar na realização dessa etapa de pré-processamento dos dados, será utilizada a linguagem de programação *Python*[1], e suas bibliotecas *Numpy*[2] e *Pandas*[3].

Inicialmente, ocorreu uma limpeza geral no conjunto de dados extraído, incluindo a remoção de linhas duplicadas e colunas consideradas desnecessárias para os objetivos desta pesquisa. A coluna "Nome do curso" foi removida, visto que todos os discente pertencem ao mesmo curso, assim como a coluna de "Forma de ingresso", visto que todos os alunos foram e são admitidos exclusivamente via sistema de seleção unificada (SISU). Além disso, foram excluídos do conjunto de dados os alunos com a situação de matrícula "trancada", pois, em teoria, eles não se enquadram nem como evadidos e nem como matrícula ativa, o que poderia impactar negativamente os resultados das predições.

[1]https://www.python.org
[2]https://numpy.org
[3]https://pandas.pydata.org

No que se refere a padronização, a coluna de "Renda per capita Famíliar" foi transformada em faixas de renda, calculadas com base no salário mínimo vigente. As faixas estabelecidas foram: 0 a 0.5 SM (renda menor ou igual a 0,5 salário mínimo), 0.5 a 1.0 SM (renda maior que 0,5 e menor ou igual a 1,0 salário mínimo), 1.0 a 1.5 SM (renda maior que 1,0 e menor ou igual a 1,5 salário mínimo), 1.5 a 2.5 SM (renda maior que 1,5 e menor ou igual a 2,5 salários mínimos), 2.5 a 3.5 SM (renda maior que 2,5 e menor ou igual a 3,5 salários mínimos), e > 3.5 SM (renda maior que 3,5 salários mínimos).

Na etapa seguinte, foram criadas duas novas colunas. A primeira identifica se o aluno reside na mesma cidade local de origem do campus ou se reside em outra localidade e outra classificando os alunos como cotistas ou não. Além disso, foi gerada a coluna "Faixa etária", que categoriza os alunos em grupos etários específicos. As faixas etárias estabelecidas foram: Menor de 20 anos, 20 a 29 anos, 30 a 39 anos, 40 a 49 anos, 50 a 59 anos e 60 anos ou mais.

Após essas tratativas, observou-se um desbalanceamento entre as classes "evadidos" e "não evadidos", com 463 alunos na classe de não evadidos e 380 na classe de evadidos. Para equalizar as classes e evitar o baixo desempenho dos modelos, foi aplicada a técnica de balanceamento *downsampling*. Essa técnica consiste na remoção de dados da classe maior, igualando a quantidade de amostras em ambas as classes.

Um conjunto de dados é considerado balanceado quando a quantidade de amostras para todas as classes é igual ou difere apenas por uma pequena porcentagem, garantindo que todas as classes estejam igualmente representadas em suas distribuições [25]. Esse equilíbrio é essencial para evitar enviesamento nos modelos, que poderiam ser influenciados pela predominância de uma classe majoritária.

3.3 Visão geral dos dados

Nesta etapa, foram realizados procedimentos para obter uma melhor compreensão dos dados trabalhados. Dentre os procedimentos realizados, destacam-se o cálculo de frequência e a geração de gráficos a partir dos dados presentes na base. Para isso, foram utilizadas as bibliotecas Matplotlib[4] e Pandas, que são amplamente usados para exploração de dados.

Inicialmente, foram criados gráficos de barras e tabelas elaboradas para cruzar variáveis, como a característica demográfica, etnia e a situação de matrícula, além disso, também foram cruzadas as informações entre faixa de renda per capita familiar e a situação matricula. Esses gráficos foram gerados com objetivo de analisar e identificar possíveis padrões nos dados, como por exemplo, como a situação matricula se comporta dentro de grupos específicos.

Além disso, os dados foram organizados em faixas de renda e categorias étnicas, visando facilitar a categorização e geração de gráficos para a análise estatística. Adicionalmente, uma tabela foi elaborada para descrever a distribuição mais detalhada dos discentes por matrícula, considerando as faixas de renda. Esses cálculos e análise servirão como suporte para a análise dos resultados apresentados neste trabalho.

3.4 Treinamento dos modelos

Nesta fase, foi efetuada a criação dos modelos a partir dos dados trabalhados na etapa de processamento de dados. Para isto foram

[4]https://matplotlib.org

utilizados algoritmos de aprendizagem de máquina focados na tarefa de classificação de dados (classificadores) a partir da biblioteca *scikit-learn*[5]. O *scikit-learn* é uma biblioteca criada em linguagem *Python* desenvolvida especificamente para aplicações prática de aprendizagem de máquina. Para otimizar os hiperparâmetros dos modelos de aprendizado de máquina utilizados na predição de evasão, foi aplicada a técnica de *GridSearch*[6]. Essa abordagem permite a busca exaustiva por combinações de parâmetros, garantido que os modelos sejam ajustados para alcançar o melhor desempenho possível.

Para a construção dos modelos de predição, foram selecionados os algoritmos de aprendizagem de máquina *Support Vector Machine* (SVM) e *Random Forest* (RF), que apresentaram melhores resultados na pesquisa realizada por [26] e também foram utilizados no estudo de [5]. Além desses, também foram incluídos os algoritmos *K-Nearest Neighbors* (KNN), *Logistic Regression* (LR), *Gradient Boosting* (GB) e *AdaBoost* (AB).

Visando um bom nível de assertividade dos modelos, foi utilizada a técnica *K-fold cross-validation* como estratégia para a geração e avaliação dos modelos. De acordo com a pesquisa descrita em [27], *K-fold cross-validation* consiste em dividir a base de dados de forma aleatória em K subconjuntos (em que K é definido previamente) com aproximadamente a mesma quantidade de amostras em cada um deles. A cada iteração, um conjunto formado por K-1 subconjuntos são utilizados para treinamento e o subconjunto restante será utilizado para teste, gerando K predições distintas. Cada um dos K conjuntos de predições serão avaliados a partir de métricas pré-estabelecidas (seção 3.5). O resultado final de cada métrica será a média aritmética dos K valores gerados para aquela métrica. A Figura 3 ilustra o funcionamento do *K-fold cross-validation* considerando K=5.

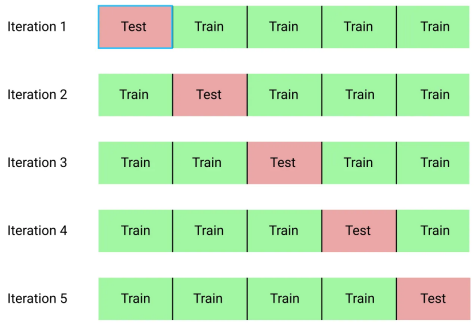

Figura 3: K-fold Cross Validation. via Rahil Shaikh. (2018).

3.5 Avaliação dos modelos

Nesta etapa, propõe-se realizar a avaliação dos modelos gerados a partir da metodologia aplicada. É importante avaliar os modelos de predição, entendendo suas qualidades e características antes de serem discutidos nos resultados. Neste sentido, após feitas as predições, estas foram categorizadas em quatro grupos a saber:

- **Verdadeiros Positivos (VP)**: casos em que o modelo previu corretamente a classe positiva.
- **Falsos Positivos (FP)**: esses são os casos em que o modelo previu incorretamente a classe positiva (porém, na verdade, era negativa).
- **Verdadeiros Negativos (VN)**: São os casos em que o modelo previu corretamente a classe negativa.
- **Falsos Negativos (FN)**: casos em que o modelo previu incorretamente a classe negativa (quando, na verdade, era positiva).

A partir desta categorização, foi construída uma matriz de confusão, cujo formato é ilustrado na Figura 6.

	NÃO EVADIU	EVADIU
NÃO EVADIU	VERDADEIRO NEGATIVO (VN)	FALSO POSITIVO (FP)
EVADIU	FALSO NEGATIVO (FN)	VERDADEIRO POSITIVO (VP)

Figura 4: Matriz de confusão. via Autor, (2024).

Após a construção da matriz de confusão, calculou-se as métricas de avaliação. A primeira métrica, acurácia, indica uma performance geral do modelo, ou seja, dentre todas as classificações, quantas o modelo classificou corretamente. A fórmula do cálculo da acurácia é exibida em (1):

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FN} + \text{VN} + \text{FP}} \tag{1}$$

A segunda métrica, precisão, indica o nível de acerto do modelo ao classificar corretamente os alunos evadidos entre todas as predições feitas para a classe "evadido".

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{2}$$

A sensibilidade, indica a capacidade do modelo de identificar corretamente os alunos evadidos dentre todos os que, de fato, evadiram.

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{3}$$

A especificidade avalia a capacidade do modelo de identificar corretamente os alunos que não evadiram, dentre todos aqueles que realmente permaneceram matriculados.

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}} \tag{4}$$

O F1-Score é uma métrica que combina a precisão e a sensibilidade em uma única medida, utilizando a média harmônica em vez

[5]https://scikit-learn.org/
[6]https://scikit-learn.org/stable/modules/grid_search.html

da média aritmética. O F1-Score é útil para avaliar como o modelo equilibra a capacidade de identificar corretamente os alunos evadidos (sensibilidade) e a precisão dessas predições.

$$\text{F1-score} = 2 * \frac{\text{precisão} * \text{sensibilidade}}{\text{precisão} + \text{sensibilidade}} \quad (5)$$

Também será gerada a curva ROC *(Receiver Operating Characteristic)* e calculada a AUC *(Area Under the Curve)*. A curva ROC é uma representação gráfica que apresenta o desempenho dos modelos, em relação a capacidade de separação referente as taxas de verdadeiros positivos e falsos positivos. Já a AUC mede a área sobre a curva ROC refletindo a capacidade do modelo de distinguir sobre as classes positivas e negativas.

O valor da AUC varia de 0 a 1. Um valor próximo de 0 indica que o modelo classifica as observações de forma totalmente errada. Uma AUC de 0.5 significa que o modelo não apresenta capacidade discriminativa, ou seja, seu desempenho é equivalente ao de um classificador aleatório. Já uma AUC próxima de 1 indica um excelente desempenho, onde o modelo consegue distinguir corretamente as classes em quase todas as situações.

4 Resultados alcançados

Como primeiros resultados, foram gerados gráficos e tabelas a partir das frequências de valores de cada variável. A Figura 5 apresenta um gráfico que cruza as informações da etnia com a situação de matrícula dos discentes. É possível verificar que brancos e pardos são a maioria dos alunos que estão cursando ou passaram pelo curso.

A Figura 6 ilustra o cruzamento das frequências das variáveis "situação de matrícula" e "faixa de renda per capita". É possível observar que o maior índice de discentes está na faixa de renda ">3.5 SM", ou seja, maior que 3,5 salários mínimos.

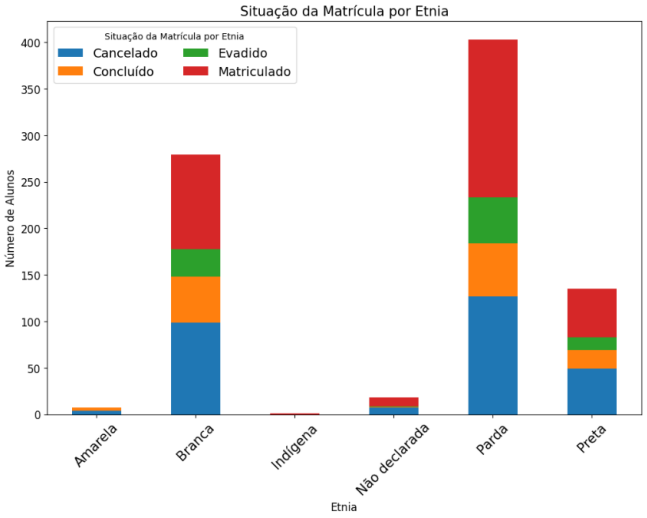

Figura 5: Gráfico Etnia x Situação matricula, via Autor (2024).

Com relação ao procedimento de balanceamento dos dados, haviam 463 discentes não evadidos e 380 alunos evadidos. Após o procedimento de *downsampling*, as quantidades de cada classe foram equalizadas, resultando em 380 discentes para a classe não evadidos e 380 para evadidos..

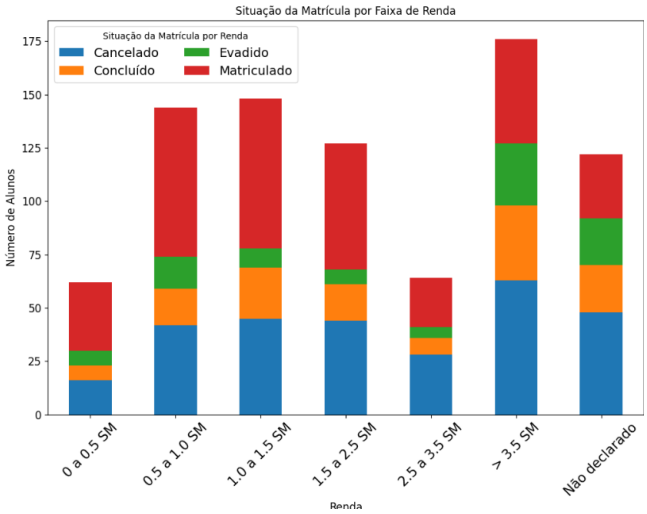

Figura 6: Gráfico Renda x Situação matricula, via Autor (2024).

Na aplicação da técnica de *k-fold cross-validation* com k=5, os dados foram divididos em 70% para treino e 30% para teste. Após o balanceamento dos dados e aplicação da técnica de *k-fold cross-validation*, os modelos foram otimizados com hiperparâmetros ajustados utilizando o *GridSearch* para melhora do desempenho, através da seleção de parâmetros. Neste contexto, o modelo *AdaBoost* utilizou o algoritmo *SAMME* com taxa de aprendizado de 0.1 e 50 estimadores. O *Gradient Boosting* foi configurado com profundidade máxima de 3, taxa de aprendizado de 0.1, mínimo de 2 amostras por folha, 10 estimadores, e semente aleatória para reprodutibilidade. O *Support Vector Machine* usou kernel RBF, regularização C=1, e cálculo de probabilidade ativado. O *Random Forest* operou com profundidade ilimitada, mínimo de 2 amostras por folha, 5 para divisão de nós e 50 árvores. O *Logistic Regression* empregou penalização L1, constante de regularização C=1, solver liblinear, e 5000 interações máximas. Por fim, o *K-Nearest Neighbors* foi ajustado com 5 vizinhos, pesos uniformes e distância Euclidiana.

Após cada uma das cinco interações do *k-fold cross validation*, os modelos foram avaliados utilizando as métricas previamente descritas na seção 3.5. A Figura 7 apresenta a matriz de confusão com os valores de erros e acertos de todos os modelos.

A partir da Figura 7, verifica-se que o SVM classificou corretamente 80 alunos como evadidos (VP) e 64 como não evadidos (VN). Já o número de falsos negativos ficou em 43 ocorrências, e a quantidade de falsos positivos foi de 41 ocorrências.

Para o *Gradient Boosting*, o modelo identificou corretamente 81 alunos evadidos (VP) e 72 não evadidos (VN). Apesar de 42 falsos negativos e 33 falsos positivos, este modelo apresentou um desempenho melhor que o modelo gerado pelo SVM, tanto pelo maior número de acertos quanto pelo menor número de erros.

Já o *AdaBoost* classificou corretamente 96 alunos evadidos (VP) e 56 não evadidos (VN). Este modelo também apresentou o maior

	NÃO EVADIU	EVADIU
NÃO EVADIU	AB: 56 GB: 72 SVM: 64 RF: 68 LR: 74 KNN: 67	AB: 49 GB: 33 SVM: 41 RF: 37 LR: 31 KNN: 38
EVADIU	AB: 27 GB: 42 SVM: 43 RF: 51 LR: 51 KNN: 57	AB: 96 GB: 81 SVM: 80 RF: 72 LR: 72 KNN: 66

Figura 7: Matriz de confusão dos modelos. via Autor (2024).

número de falsos positivos - 49 ocorrências, e o menor número de falsos negativos - 27 ocorrências dentre todos os modelos, apresentando-se como um bom modelo para predizer alunos que tendem de fato a evadir.

Os modelos *Random Forest* e *Logistic Regression* apresentaram desempenhos semelhantes, classificando corretamente 72 alunos evadidos (VP). O RF acertou 68 não evadidos (VN) com 51 falsos negativos e 37 falsos positivos, enquanto o LR teve 74 acertos de não evadidos, com 51 falsos negativos e o menor número de falsos positivos (31). Já o modelo *K-Nearest Neighbors* foi o mais conservador, com 66 acertos de evadidos e 67 de não evadidos, apresentando 57 falsos negativos e 38 falsos positivos, o que indica que esse modelo apresentou menor capacidade de prever corretamente os alunos que evadiram, quando comparado com os demais.

Considerando os dados presentes na matriz de confusão, a Tabela 3 exibe os níveis das métricas de acurácia, precisão, sensibilidade, medida F1, e especificidade de cada modelo gerado.

Tabela 3: Resultados dos Algoritmos

Algoritmo	Acc	Prec	Sens	F1	Esp
KNN	0.6013	0.6029	0.5974	0.5994	0.6053
LR	0.6395	0.6429	0.6421	0.6404	0.6368
RF	0.6447	0.6322	0.6947	0.6615	0.5947
SVM	0.6329	0.6192	0.7053	0.6579	0.5605
GB	0.6303	0.6014	0.7868	0.6782	0.4737
AB	0.6500	0.6029	0.8868	0.7173	0.4132

Acc: Acurácia, **Prec**: Precisão, **Sens**: Sensibilidade, **F1**: F1-score, **Esp**: Especificidade, **SVM**: Support Vector Machine, **RF**: Random Forest, **KNN**: K-Nearest Neighbors , **LR**: Logistic Regression, **GB**: Gradient Boosting, **AB**: AdaBoost.

O modelo com melhor desempenho geral, baseado na métrica *F1-score*, foi o *AdaBoost*, alcançando um valor de 0.7173, demonstrando um bom equilíbrio entre precisão e sensibilidade. Em segundo lugar, o *Gradient Boosting* obteve um valor de 0.6782. Em terceiro lugar, o *Random Forest* obteve 0.6615, acompanhado de perto pelo *Support Vector Machine*, com 0.6579. O *Logistc Regression* apresentou um *F1-score* de 0.6404, enquanto o *K-Nearest Neighbors* registrou o menor desempenho, com 0.5994.

Considerando o objetivo da pesquisa, a análise da taxa de sensibilidade é especialmente relevante, pois reflete a capacidade do modelo de identificar corretamente os alunos propensos à evasão. Prever corretamente esses alunos é importante porque a gestão da instituição pode usar essas informações para intervir proativamente, implementando medidas para prevenir a evasão antes que ela ocorra. Por exemplo, pode-se oferecer suporte acadêmico, psicossocial ou apoio financeiro por meio de auxílios. O modelo gerado pelo AdaBoost obteve o maior nível de sensibilidade, atingindo uma taxa de 0,8868, enquanto o Gradient Boosting também apresentou bons resultados, alcançando 0,7868. No entanto, ao avaliar a baixa especificidade desses modelos, observa-se que eles tendem a cometer mais falsos positivos, ou seja, identificar alunos como propensos à evasão quando, na realidade, não o são, o que pode levar a intervenções desnecessárias por parte da instituição. Por outro lado, os algoritmos SVM e *Random Forest* apresentaram um desempenho mais equilibrado, com SVM alcançando uma taxa de sensibilidade de 0.7053 e *Random Forest* 0.6947. Esses modelos também obtiveram uma especificidade mais alta, indicando uma maior taxa de acertos em relação aos alunos que não evadiram.

A Figura 8 apresenta o desempenho dos modelos em relação à curva AUC-ROC. Observa-se que o modelo AdaBoost apresenta um valor mais próximo de 1, indicando um bom desempenho na taxa de acertos. Em contrapartida, o KNN não obteve um bom desempenho, conforme refletido nas métricas gerais. O modelo gerado pelo algoritmo Random Forest se destacou por apresentar um equilíbrio mais adequado entre precisão, sensibilidade e especificidade. Contudo, seu desempenho ainda é inferior ao apresentado pelo AdaBoost.

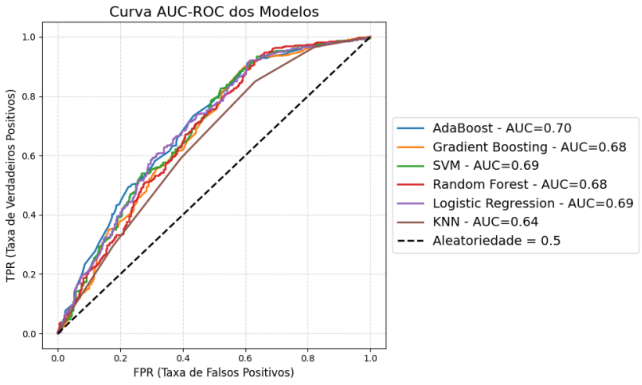

Figura 8: Curva AUC-ROC. via Autor (2024).

O Figura 9 destaca os fatores mais relevantes na previsão de evasão acadêmica. O Ano de Ingresso se sobressai como a variável mais significativa, sugerindo que mudanças institucionais e contextos específicos de cada ano impactam diretamente a permanência dos alunos. Isso pode estar relacionado ao ambiente acadêmico no momento da admissão e ao tempo necessário para conclusão do curso, influenciado por reprovações e trancamentos.

Fatores como Faixa Etária, Etnia e Localização também exercem influência, reforçando o impacto de aspectos demográficos e geográficos. Sexo e condição de Cotista, apesar de menor impacto, ainda contribuem para o modelo. Esses achados reforçam a necessidade de políticas de inclusão, suporte financeiro e estratégias que favoreçam a adaptação dos alunos ao ambiente acadêmico.

Com base no objetivo do trabalho, o *AdaBoost* demonstrou o melhor desempenho em classificar corretamente o número de alunos evadidos. No entanto, é importante destacar que, para manter um

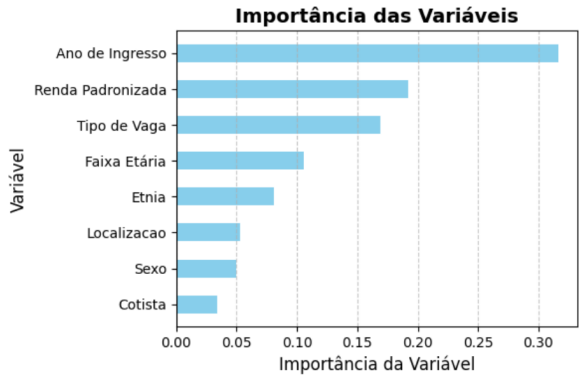

Figura 9: Nível de importância das variáveis. via Autor (2024).

equilíbrio entre a identificação de evadidos e não evadidos, o Gradient Boosting é uma boa opção, contribuindo significativamente para a predição de alunos propensos a evadir.

5 Conclusões e trabalhos futuros

Este trabalho se dedicou à análise de dados de um curso em uma instituição federal, adotando uma abordagem centrada na criação de modelos preditivos. Para a construção desses modelos, foram selecionados algoritmos que são amplamente reconhecidos na literatura acadêmica por sua eficácia e robustez na predição de evasão escolar. Esses algoritmos foram escolhidos com base em suas boas taxas de desempenho em estudos anteriores, o que garante uma base sólida para a análise. Para validar os modelos gerados, foram empregadas diversas métricas de avaliação, incluindo acurácia, precisão, sensibilidade, especificidade, F1-score e a curva AUC-ROC. Essas métricas são fundamentais para verificar e garantir que os modelos estejam sendo assertivos. A utilização das diferentes métricas utilizadas neste trabalho proporciona uma visão abrangente do desempenho dos modelos, assegurando que as intervenções propostas sejam fundamentadas em resultados mensuráveis.

A análise deste trabalho aponta para diversas possibilidades de continuidade na pesquisa sobre a predição de evasão escolar em cursos de tecnologia. Uma direção promissora é a expansão da base de dados, incluindo informações adicionais como dados de desempenho acadêmico, além da coleta de dados ao longo do tempo para uma visão sobre a tendência dos níveis de de evasão do curso. Além disso, a implementação do modelo preditivo como solução educacional deve ser avaliada em termos de suas implicações práticas e eficácia no mundo real. Isso pode incluir o desenvolvimento de ferramentas visuais e analíticas para auxiliar gestores e educadores na tomada de decisões para combater a evasão e promover o sucesso dos alunos.

References

[1] Censo de Educação Superior. Painel estatístico, 2024. URL https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados.

[2] Instituto SEMESP. Mapa do ensino superior no brasil 2024, 2024. URL https://www.semesp.org.br/wp-content/uploads/2024/04/mapa-do-ensino-superior-no-brasil-2024.pdf.

[3] Ramesh Sharda, Dursun Delen, and Efraim Turban. *Business Intelligence e Análise de Dados para Gestão do Negócio-4.* Bookman Editora, 2019.

[4] Patricia S. Abril and Robert Plant. Bhinder, bhavneet and gilvary, coryandar and madhukar, neel s and elemento, olivie. *Cancer discovery*, 11(4):900–915, 2021.

[5] Ronei dos Santos Oliveira and Francisco Petronio Alencar de Medeiros. Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação. *Revista Brasileira de Informática na Educação*, 3(1):4, jan 2024. doi: 10.5753/rbie.2024.3542. URL https://journals-sol.sbc.org.br/index.php/rbie/article/view/3542.

[6] Ronei dos Santos Oliveira and Francisco Petronio Alencar de Medeiros. Inteligência artificial na educação: aplicações do aprendizado de máquina para apoiar a aprendizagem adaptativa. *Revista Multidisciplinar do Vale do Jequitinhonha-ReviVale*, 1(1), 2021.

[7] Henrique Rosario Carvalho Esteves, Carlos Alberto Dias, Ciro Meneses Santos, and Agnaldo Keiti Higuchi. Evasão escolar no ensino superior: uma revisão literária entre os anos de 2014 a 2020. *Research, Society and Development*, 10(3), 2021.

[8] Chaiane de Medeiros Rosa. Limites da democratização da educação superior: entraves na permanência e a evasão na universidade federal de goiás. *Poíesis Pedagógica*, 12(1):240–257, 2014.

[9] Cidmar Ortiz dos Santos, Luiz Alberto Pilatti, and Roberto Bondarik. Evasão no ensino superior brasileiro: conceito, mensuração, causas e consequências. *Debates em Educação*, 14(35):294–314, 2022.

[10] Adriana Cristina Kozelski and Silvana Hammerschmidt. Políticas públicas: Recurso ou solução para evasão universitária? *Revista on line de Política e Gestão Educacional*, (6):31–40, 2009.

[11] ML Boero and AF de SILVA. A evasão escolar em uma universidade privada. In *CONGRESSO BRASILEIRO DE EDUCAÇÃO EM ENGENHARIA*, volume 34, 2006.

[12] Armênia Chaves Fernandes VIEIRA, Erica de Lima GALLINDO, and Hobson Almeida CRUZ. Plano estratégico para permanência e êxito dos estudantes do ifce. *Fortaleza: IFCE*, 2017.

[13] Luciana Soares Rodrigues, Tarcísio Laerte Gontijo, Ricardo Bezerra Cavalcante, Patrícia Peres de Oliveira, and Sebastião Júnior Henrique Duarte. A evasão em um curso de especialização em gestão em saúde na modalidade a distância. *Interface-Comunicação, Saúde, Educação*, 22:889–901, 2018.

[14] Marillia Gabriella Duarte Fialho et al. A evasão escolar e a gestão universitária: o caso da universidade federal da paraíba. 2014.

[15] Margareth Nunes Silva and Maria Aparecida Monteiro da Silva. Evasão escolar e educação profissional técnica brasileira um estudo de revisão. *RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218*, 4(12):e4124707–e4124707, 2023.

[16] Katti Faceli, Ana Carolina Lorena, João Gama, and André Carlos Ponce de Leon Ferreira de Carvalho. *Inteligência artificial: uma abordagem de aprendizado de máquina.* LTC, 2011.

[17] Arthur L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[18] Pedro Domingos. *O algoritmo mestre: como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo.* Novatec Editora, 2017.

[19] Gabriela Miana de Mattos Paixão, Bruno Campos Santos, Rodrigo Martins de Araujo, Manoel Horta Ribeiro, Jermana Lopes de Moraes, and Antonio L Ribeiro. Machine learning na medicina: Revisão e aplicabilidade. *Arquivos brasileiros de cardiologia*, 118:95–102, 2022.

[20] Marcos Kalinowski, Tatiana Escovedo, Hugo Villamizar, and Hélio Lopes. *Engenharia de Software para Ciência de Dados: Um guia de boas práticas com ênfase na construção de sistemas de Machine Learning em Python.* Casa do Código, 2023.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, 2002.

[22] Sotiris B Kotsiantis, CJ Pierrakeas, and Panayiotis E Pintelas. Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems: 7th International Conference, KES 2003, Oxford, UK, September 2003. Proceedings, Part II 7*, pages 267–274. Springer, 2003.

[23] Jorge Ramos, Rodrigo Rodrigues, João Silva, and Pamella Oliveira. CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, (1):1092–1101, 2020. doi: 10.5753/cbie.sbie.2020.1092. URL https://sol.sbc.org.br/index.php/sbie/article/view/12865.

[24] Cristiane Aparecida dos Santos Baggi and Doraci Alves Lopes. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 16(02):355–374, 2011.

[25] Camila Maione et al. Balanceamento de dados com base em oversampling em dados transformados, 2020.

[26] Vitor Hugo Barbosa dos Santos, Daniel Victor Saraiva, and Carina Teixeira de Oliveira. Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1196–1210. SBC, 2021.

[27] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.