Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 20/Nov/2024; 1st round notif.: 06/Mar/2025; New version: 04/Apr/2025; 2nd round notif.: 19/May/2025;
Camera ready: 24/May/2025; Edition review: 17/Sep/2025; Available online: 17/Sep/2025; Published: 17/Sep/2025;

Previsão do risco de evasão universitária utilizando aprendizado de máquina: Características e dinâmica da evasão nos cursos de Engenharia da UFRJ-Macaé

Title: Predicting the risk of university dropout using machine learning: Characteristics and dynamics of dropout in Engineering courses at UFRJ-Macaé

Título: Predecir el riesgo de abandono universitario utilizando el aprendizaje automático: Características y dinámica del abandono en las carreras de Ingeniería de la UFRJ-Macaé

Pedro Andrietta Chagas
Universidade Federal do Rio de Janeiro
ORCID: 0009-0003-3380-9666
pchagas96@gmail.com

Janaina Sant'Anna Gomide
Universidade Federal do Rio de Janeiro
ORCID: 0000-0003-3310-4669
janainagomide@macae.ufrj.br

Laura Emmanuella Alves dos Santos Santana
Universidade Federal do Rio de Janeiro
ORCID: 0000-0003-2086-3471
lauraemmanuella@macae.ufrj.br

Resumo

O fenômeno da evasão universitária tem sido amplamente investigado devido aos prejuízos em recursos financeiros e em mão de obra, por isso, entender sua dinâmica é importante para identificar soluções e estratégias preventivas. Este estudo emprega técnicas de ciência de dados e aprendizado de máquina para prever o risco de evasão universitária e analisar as características e dinâmicas da evasão nos cursos de Engenharia da Universidade Federal do Rio de Janeiro na cidade de Macaé, Brasil. Utilizando a metodologia CRISP-DM, foram treinados modelos de Regressão Logística, Árvore de Decisão e eXtreme Gradient Boosting (XGBoost), sendo otimizados por meio da otimização bayesiana. Conforme os resultados dos modelos, observou-se que o desempenho acadêmico dos alunos no seu primeiro período é suficiente para identificar, com AUC de 0,80, os alunos evadidos do conjunto de testes. Na análise exploratória dos dados também foi revelado que mais alunos evadiram do que se formaram, sendo que 75% das desistências ocorreram nos três primeiros semestres. Além disso, observou-se uma forte correlação entre o baixo desempenho acadêmico e o risco de evasão, com destaque para as disciplinas de Cálculo e Física. Este trabalho evidencia como os dados de uma universidade podem ser explorados para identificar padrões e tendências no comportamento dos alunos e como o aprendizado de máquina pode ser empregado como ferramenta estatística para extrair informações valiosas de grandes volumes de dados, auxiliando na melhoria da qualidade e acessibilidade da educação no ensino superior.

Palavras-chave: *Evasão; Universidade; Engenharia; Desempenho acadêmico; Aprendizado de máquina.*

Abstract

The phenomenon of university dropout has been widely investigated due to losses in financial resources and labor, so understanding its dynamics is important to identify solutions and preventive strategies. This study uses data science and machine learning techniques to predict the risk of university dropout and analyze the characteristics and dynamics of dropout in Engineering courses at the Federal University of Rio de Janeiro in the city of Macaé, Brazil. Using the CRISP-DM methodology, Logistic Regression, Decision Tree, and eXtreme Gradient Boosting (XGBoost)

Cite as: Chagas, P. A., Gomide, J. S. & Santana, L. E. A. S. (2025). Previsão do risco de evasão universitária utilizando aprendizado de máquina: Características e dinâmica da evasão nos cursos de Engenharia da UFRJ-Macaé. Revista Brasileira de Informática na Educação, vol, 1226-1247. https://doi.org/10.5753/rbie.2025.5196.

models were trained, being optimized through Bayesian optimization. According to the results of the models, it was observed that the academic performance of students in their first period is sufficient to identify, with an AUC of 0.80, students who dropped out of the test set. In the exploratory analysis of the data, it was also revealed that more students dropped out than graduated, with 75% of dropouts occurring in the first three semesters. Furthermore, a strong correlation was observed between low academic performance and the risk of dropping out, particularly in the subjects of Calculus and Physics. This work highlights how a university's data can be explored to identify patterns and trends in student behavior and how machine learning can be used as a statistical tool to extract valuable information from large volumes of data, helping to improve quality and accessibility of education in higher education.
Keywords: *Evasion; University; Engineering; Academic performance; Machine learning.*

Resumen

El fenómeno de la deserción universitaria ha sido ampliamente investigado debido a las pérdidas de recursos financieros y laborales, por lo que comprender su dinámica es importante para identificar soluciones y estrategias preventivas. Este estudio utiliza técnicas de ciencia de datos y aprendizaje automático para predecir el riesgo de deserción universitaria y analizar las características y dinámicas de la deserción en carreras de Ingeniería de la Universidad Federal de Río de Janeiro en la ciudad de Macaé, Brasil. Utilizando la metodología CRISP-DM se entrenaron los modelos de Regresión Logística, Árbol de Decisión y Aumento de gradiente extremo (XGBoost), siendo optimizados mediante optimización bayesiana. Según los resultados de los modelos, se observó que el rendimiento académico de los estudiantes en su primer período es suficiente para identificar, con un AUC de 0,80, a los estudiantes que abandonaron el conjunto de pruebas. En el análisis exploratorio de los datos, también se reveló que más estudiantes desertaron de los que se graduaron, ocurriendo el 75% de los abandonos en los primeros tres semestres. Además, se observó una fuerte correlación entre el bajo rendimiento académico y el riesgo de abandono, particularmente en las materias de Cálculo y Física. Este trabajo destaca cómo se pueden explorar los datos de una universidad para identificar patrones y tendencias en el comportamiento de los estudiantes y cómo el aprendizaje automático se puede utilizar como herramienta estadística para extraer información valiosa de grandes volúmenes de datos, ayudando a mejorar la calidad y la accesibilidad de la educación en la educación superior.
Palabras clave: *Evasión; Universidad; Ingeniería; Rendimiento académico; Aprendizaje automático.*

1 Introdução

A evasão estudantil nas universidades brasileiras e mundiais é uma questão complexa e multifacetada que afeta o sistema educacional como um todo. A evasão ocorre quando os estudantes abandonam seus cursos antes de concluí-los, seja por motivos pessoais, acadêmicos, financeiros ou sociais. Esse fenômeno tem sido motivo de preocupação tanto para as instituições de ensino quanto para os governos, uma vez que impacta negativamente a qualificação profissional da população (Jesus & Gusmão, 2024; Oliveira & Medeiros, 2024; Teodoro & Kappel, 2020).

No contexto brasileiro, a evasão estudantil é um problema persistente. Muitos estudantes entram na universidade com grandes expectativas, mas acabam desistindo ao longo do caminho. Dentre as principais razões para a evasão estão a falta de recursos financeiros para se sustentar durante os estudos, a necessidade de trabalhar para ajudar a família, a falta de suporte acadêmico, a dificuldade de adaptação à vida universitária e a falta de motivação ou interesse pelo curso escolhido (Coimbra et al., 2021). Além dos fatores individuais, a evasão também está relacionada a problemas estruturais do sistema educacional, inclusive da educação básica (Carvalhaes et al., 2022), uma vez que este sistema enfrenta desafios como a falta de infraestrutura adequada e a falta de professores qualificados e motivados, fazendo com que os estudantes cheguem à universidade com baixo preparo para acompanhar o currículo universitário.

Além disso, a pandemia da COVID-19 desencadeou uma grave situação econômica no país, incluindo a área da educação, desde o ensino básico até o superior, restringindo o acesso aos campus universitários e gerando um desafio para garantir a continuidade do processo de ensino-aprendizado, agravando o problema da evasão estudantil (Alharbi, 2020; Crawford et al., 2020; Ebner et al., 2020; Murphy, 2020; Regehr & Goel, 2020). Segundo o Mapa do Ensino Superior, publicado pelo Instituto Semesp, o ano de 2020 registrou os maiores índices de evasão de alunos do ensino superior no Brasil de toda a série histórica, uma taxa de 28,5% de evasão nos cursos presenciais e 33,5% nos cursos à distância (Semesp, 2022). Esses dados corroboram os resultados do trabalho de Klitzke e Carvalhaes (2023), que, ao analisar os fatores associados à evasão de curso na Universidade Federal do Rio de Janeiro para a coorte de 2014.1, encontrou que 33% destes alunos evadiram do seu curso até o 6º período. Levando em conta apenas esses dados e o custo anual médio de um aluno para uma instituição superior (Bielschowsky & Amaral, 2022), é possível entender a existência de um prejuízo financeiro considerável por ano para o Estado que já sofre com o impacto econômico e social dessa tendência, pela menor produção de mão-de-obra qualificada e menor produção científica (Lobo, 2012).

A exploração de dados educacionais com o intuito de melhorar a experiência de aprendizado e a eficiência da gestão das instituições de ensino vem sendo utilizada desde 1979, quando a Open University (Reino Unido) já podia analisar mais de 10 anos de progresso de seus alunos a distância, dando origem ao campo do Learning Analytics (LA). Desde então, com a emergência do ensino à distância (EAD) e dos conceitos de mineração de dados e big data, o LA tem sido pauta das discussões educacionais, e a maioria das iniciativas que o empregam envolvem esforços para melhorar as taxas de retenção dos alunos ou criar uma experiência de ensino personalizada (Hernández-de-Menéndez et al., 2022).

Na sua revisão sistemática, de Oliveira et al. (2021) argumentam que a atual transformação digital oferece às universidades uma variedade de oportunidades e facilidades. Nesse novo

contexto, essas instituições podem fazer uso apropriado dos dados produzidos pela interação dos alunos com a universidade, bem como compreender como diferentes aspectos variam em decorrência de determinadas intervenções. Os autores mostram como a maioria dos trabalhos tenta prever o status do aluno para tentar evitar sua evasão da instituição. Quanto aos dados utilizados para tais previsões, os dados dos estudantes (informações pessoais e acadêmicas) são a categoria de dados mais utilizada. No que diz respeito à tarefa de aprendizado de máquina utilizada para a previsão, a maioria dos estudos analisados utilizou técnicas de classificação, com vários algoritmos de classificação diferentes. A técnica de validação mais utilizada é a validação cruzada, e as métricas de avaliação mais frequentes são aquelas baseadas na matriz de confusão. Alguns trabalhos utilizando dados de universidades brasileiras podem ser encontrados em: Gomes (2021), Manhães et al. (2012), Oliveira e Medeiros (2024), Oliveira Júnior (2015), Silva e Adeodato (2012) e Teodoro e Kappel (2020). Além disso, é importante destacar as questões de privacidade. Como alguns dos dados utilizados para previsões podem ser pessoais e sensíveis, a privacidade precisa ser considerada; esta preocupação também é referida em alguns dos estudos analisados.

Nesta pesquisa, o objetivo é analisar as características da evasão universitária nos cursos de Engenharia da UFRJ Macaé e aplicar técnicas de aprendizado de máquina para identificar os alunos em risco de evasão, compreendendo melhor o fenômeno do abandono. É interessante comentar que o trabalho de da Silva (2023) analisou a evasão nos cursos de Engenharia da UFRJ Macaé pela perspectiva da experiência do usuário (UX). A partir de um estudo de caso qualitativo com questionários aplicados a estudantes do 1º ao 3º período, foram identificados fatores como dificuldades financeiras, de transporte, adaptação e gestão de tempo, além de mau desempenho e defasagens de conhecimento. Enquanto o estudo de da Silva (2023) utiliza uma perspectiva qualitativa centrada na experiência do usuário (UX), identificando fatores subjetivos, o presente trabalho aplica técnicas quantitativas de ciência de dados e aprendizado de máquina para prever a evasão e explorar padrões em grandes volumes de dados acadêmicos.

2 Delimitação do escopo e limitações do estudo

A análise realizada neste estudo se baseia exclusivamente em dados de desempenho acadêmico, devido a restrições institucionais no acesso a informações de natureza socioeconômica, cultural e comportamental dos alunos. A obtenção desses dados demandaria aprovações específicas relacionadas à privacidade e proteção de informações sensíveis, o que não foi viável no contexto desta pesquisa. Portanto, a modelagem proposta busca avaliar especificamente a relação entre o desempenho acadêmico e a evasão universitária, reconhecendo que outros fatores podem exercer influência sobre esse fenômeno.

Apesar dessa limitação em relação aos dados utilizados, é importante chamar a atenção para estudos anteriores que apontaram que o desempenho acadêmico, notadamente, em disciplinas de matemática e ciências naturais, tem forte relação com a evasão nos cursos de engenharia. Godoy e Almeida (2017) comentam em seu trabalho que "as disciplinas da área de Matemática e das Ciências Naturais atreladas à fraca formação, nessas mesmas áreas, na Educação Básica contribuem, consideravelmente, para a evasão". Já Christo et al. (2018) diz que os resultados de sua pesquisa "demonstram que o fator financeiro não é o principal no ato da desistência, pois a

maior parte dos alunos desistentes (61%) declarou desistir por motivos acadêmicos e apenas 12% por motivos socioeconômicos".

Além disso, embora este estudo se concentre exclusivamente em cursos de Engenharia, essa delimitação é justificada pela alta taxa de evasão observada historicamente nesse setor no Brasil (Christo et al., 2018; Godoy & Almeida, 2017). A evasão nos cursos de Engenharia representa um desafio significativo para o país, pois impacta diretamente na formação de profissionais essenciais para setores estratégicos do desenvolvimento nacional, como infraestrutura, tecnologia e inovação. Ademais, considerando que os fatores associados à evasão podem variar entre diferentes áreas do conhecimento, as análises setoriais como esta tornam-se fundamentais para a construção de estratégias específicas de retenção de alunos. Assim, os resultados aqui apresentados contribuem para um melhor entendimento desse fenômeno em um setor crítico e podem servir como base para estudos futuros em áreas afins no ensino superior.

3 Trabalhos Relacionados

O trabalho de Jesus e Gusmão (2024) apresenta uma revisão sistemática sobre métodos de mineração de dados e aprendizado de máquina para prever e mitigar a evasão estudantil, destacando o uso de algoritmos baseados em árvores de decisão e a predominância de estudos sobre ensino presencial. O estudo enfatiza a necessidade de dados diversificados e modelos ajustados ao contexto educacional brasileiro. Já em Oliveira e Medeiros (2024), um modelo preditivo foi desenvolvido para identificar estudantes em risco de evasão em cursos de graduação a partir de dados de autoavaliação dos próprios alunos, com acurácia de 87,97%, enfatizando variáveis como satisfação com o curso e apoio institucional. O estudo destaca a importância de variáveis como satisfação com o curso e apoio institucional, contribuindo para estratégias educacionais de retenção e apoio ao estudante.

Teodoro e Kappel (2020) exploraram algoritmos aplicados a dados de universidades públicas brasileiras obtidos do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), com a Random Forest alcançando 80% de acerto na previsão de evasão. Em Oliveira Júnior (2015), foram empregadas técnicas de mineração de dados, incluindo a seleção de subconjuntos de atributos e a criação de novos atributos, para identificar padrões de evasão na UTFPR, contribuindo para análises mais aprofundadas do problema e possibilitando a criação de ferramentas de visualização em tempo real. Já Silva e Adeodato (2012) analisaram dados de desempenho acadêmico de oito cursos da Universidade Federal do Pernambuco (UFPE) e propuseram modelos baseados em regressão, destacando a importância de fatores socioeconômicos, culturais e comportamentais.

Fora do contexto brasileiro, Aulck et al. (2019) analisaram dados do primeiro ano de 66.060 estudantes nos EUA, demonstrando que informações de desempenho acadêmico inicial são mais úteis que dados demográficos para prever evasão (AUC = 0,811). De forma semelhante, Nagy e Molontay (2018) exploraram dados pré-admissão e do primeiro ano de 15.825 estudantes em Budapeste, com redes neurais e Gradient Boosted Trees apresentando os melhores resultados (AUC = 0,79 e 0,776, respectivamente). Ambos destacam a relevância de previsões precoces no combate à evasão.

Comparando os artigos citados com o presente trabalho, é possível identificar semelhanças e diferenças nos objetivos, acesso aos dados, número de cursos estudados e resultados obtidos. Embora todos compartilhem o objetivo comum de prever a evasão no ensino superior, eles diferem na utilização de dados e técnicas de aprendizado de máquina. Enquanto este trabalho utiliza apenas dados dos alunos de Engenharia, os trabalhos citados empregam dados de diversos cursos. Além disso, assim como em Aulck et al. (2019) e Nagy e Molontay (2018), este estudo também se concentra no desempenho acadêmico dos alunos no primeiro período de seus cursos para prever o risco de evasão. Em contraste, alguns dos artigos citados utilizam dados socioeconômicos e demográficos, além de dados de desempenho acadêmico, enquanto outros empregam técnicas de mineração de dados, como seleção de subconjunto de atributos e criação de atributos mais complexos.

Em termos de resultados, o presente estudo obteve um AUC de 0,80, o que está em linha com o desempenho dos modelos encontrados na literatura. Esses resultados corroboram as conclusões de Nagy e Molontay (2018), que sugerem que os dados relativos ao desempenho acadêmico são mais importantes do que os dados pré-admissão e que esses são suficientes para prever a evasão com sucesso. Entretanto, assim como evidenciado por Jesus e Gusmão (2024), ter acesso a uma maior diversidade de dados de alunos também foi um desafio enfrentado neste trabalho, de forma que não foi possível seguir as recomendações de Silva e Adeodato (2012) quanto à inclusão de outros tipos de dados não ligados ao desempenho acadêmico. Finalmente, assim como sugerido por Oliveira Júnior (2015), este trabalho também aponta direções para a continuação do trabalho que incluem a implantação dos modelos na infraestrutura da universidade para permitir o monitoramento das situações dos alunos em tempo real.

4 Método

Para este trabalho, foi implementada a metodologia CRISP-DM (Cross Industry Standard Process for Data Mining), criada há mais de 20 anos, visando auxiliar na estruturação dos projetos de mineração de dados (Chapman, 2000). Para isto, foram seguidas as 5 etapas apresentadas a seguir:

4.1 Etapa 1: Compreensão do negócio

A fase de compreensão do negócio consiste em entender a problemática, investigar soluções de problemas similares na literatura e propor abordagens para resolver o problema identificado dentro das restrições do contexto em que o projeto se insere. A análise dos trabalhos relacionados revelou que o problema da evasão é multifatorial e que as abordagens de desenvolvimento de modelos de aprendizado de máquina diferem consideravelmente em termos dos atributos empregados. Desejava-se, portanto, obter o máximo de atributos possíveis para uma caracterização mais completa do fenômeno, incluindo dados socioeconômicos, pré-admissão e desempenho acadêmico. Entretanto, devido às restrições ao acesso de dados sensíveis de alunos, foram obtidos apenas dados relativos ao desempenho dos alunos no processo de admissão à universidade e no decorrer da graduação. Esta limitação de acesso a certos tipos de dados efetivamente delimita o escopo e os objetivos específicos da análise desenvolvida. Dessa forma, o presente trabalho analisa, especificamente, a influência do desempenho acadêmico dos alunos no risco de evasão e a

modelização do fenômeno da evasão a partir do mesmo.

4.2 Etapa 2: Compreensão dos dados

Esta etapa permite que os pesquisadores se familiarizem com os dados que serão utilizados. Ela envolve vários sub-passos: coleta, descrição, exploração e verificação da qualidade dos dados. Esses passos são importantes para garantir que os dados sejam adequados para uso na pesquisa e que quaisquer questões ou tendências potenciais sejam identificadas e tratadas.

4.2.1 Coleta dos dados

Foram coletados dados dos alunos dos cursos de Engenharia de Produção, Engenharia Mecânica e Engenharia Civil do Instituto Politécnico do Centro Multidisciplinar UFRJ - Macaé disponíveis no Sistema Integrado de Gestão Acadêmica (SIGA) da Universidade Federal do Rio de Janeiro. Os dados obtidos contemplam informações sobre o desempenho em cada disciplina cursada por cada um dos alunos, assim como as notas de admissão à universidade.

4.2.2 Descrição dos dados

O conjunto de dados inicial contém 56659 linhas referentes aos resultados dos alunos em cada disciplina cursada, compreendendo um total de 1452 alunos ao longo dos 11 anos de existência dos cursos de engenharia do CM UFRJ - Macaé, de 2011 a 2022. A Tabela 1 apresenta a descrição dos atributos do conjunto de dados.

Tabela 1: Descrição dos atributos do conjunto de dados.

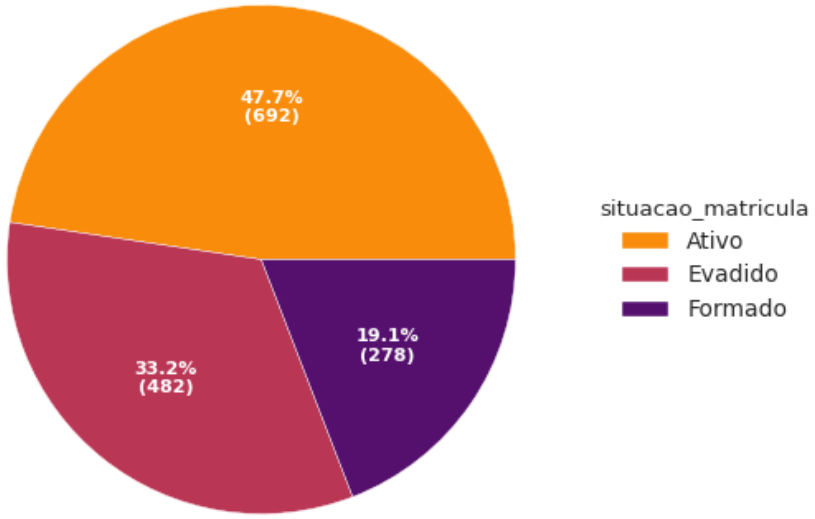
Atributo	**Descrição**	**Tipo**
primeiro_per	Ano e semestre de ingresso	Categórico
nota_humanas	Nota da prova de Ciências Humanas do ENEM	Numérico contínuo
nota_linguagem	Nota da prova de Linguagem do ENEM	Numérico contínuo
nota_matematica	Nota da prova de Matemática do ENEM	Numérico contínuo
nota_natureza	Nota da prova de Ciências da Natureza do ENEM	Numérico contínuo
nota_redacao	Nota da prova de Redação do ENEM	Numérico contínuo
grau	Nota final de cada aluno em cada disciplina cursada	Numérico contínuo
periodo	Ano e semestre em que cada disciplina foi cursada	Categórico
disciplina	Código da disciplina	Categórico
resultado	Resultado final na disciplina	Booleano
cr	Coeficiente de rendimento do aluno em cada período	Numérico contínuo
cr_acumulado	Coeficiente de rendimento total do aluno até determinado período	Numérico contínuo
curso	Curso no qual o aluno está matriculado	Categórico
situacao_matricula	Situação atual da matrícula do aluno	Categórico

4.2.3 Análise Exploratória dos dados

Em seguida, foi feita uma exploração dos dados utilizando a biblioteca Pandas, que permite a manipulação de estruturas de dados tabulados. Para visualização de padrões e inspeção dos dados, foram utilizadas as bibliotecas Matplotlib e Seaborn, que permitem a construção de gráficos e outras visualizações.

O gráfico de setores da Figura 1 evidencia a quantidade e proporção das situações da matrícula dos alunos. Nota-se que uma minoria de 278 alunos se formou, enquanto 482 alunos evadiram de seus cursos, uma quantidade 73% maior do que o número de formados.

Figura 1: Quantidade de alunos por situação.

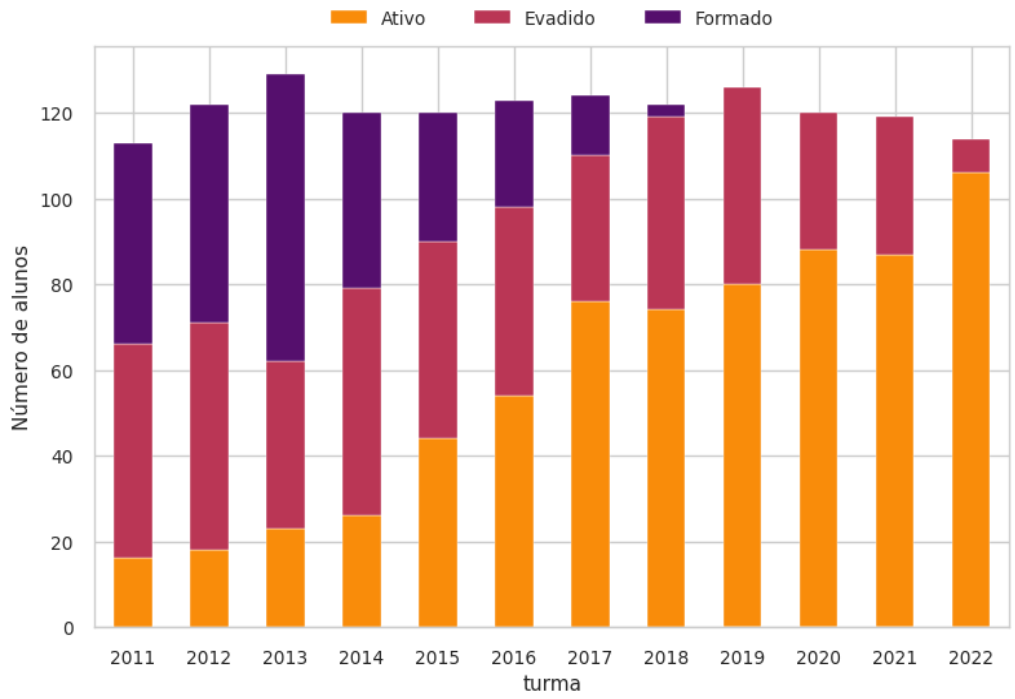


A análise da situação da matrícula por turma (ver Figura 2) revela um cenário ainda mais preocupante. A maioria dos estudantes nas turmas iniciais abandonou o curso, exceto para a turma de 2013, que tem a maior proporção de formados. Além disso, pode-se observar que uma grande parte dos estudantes das turmas iniciais ainda está ativa, mesmo após ultrapassarem significativamente o limite de tempo de 5 anos para conclusão do curso. As turmas de 2019 a 2022 não possuem formados ainda, pois não completaram o ciclo de 5 anos de curso.

Além de observar quantos alunos evadiram, é interessante também analisar o número de períodos cursados pelos alunos até o momento da análise. Para isso, a partir da contagem dos coeficientes de rendimento (CR) em cada período para cada aluno, criou-se o atributo *periodos_cursados*. O gráfico boxplot da Figura 3 ilustra a distribuição desse atributo para cada situação de matrícula.

Essa análise é importante, pois permite identificar padrões no desempenho acadêmico dos alunos, ajudando a universidade a entender melhor as necessidades dos estudantes e aprimorar sua oferta educacional. Além disso, observar em qual momento do curso os alunos costumam evadir pode ser útil para uma eventual realocação de vagas na universidade ou abertura de chamadas para outros alunos ingressarem. Nota-se primeiramente que a distribuição do número de períodos

Figura 2: Quantidade de alunos por situação de cada turma.



cursados dos alunos formados é simétrica em torno de 10, o que é de se esperar, visto que a expectativa de tempo de curso é de 10 períodos. Os alunos que se formam em menos tempo do que seria teoricamente possível (7 períodos ou menos) são aqueles que ingressaram por transferência de instituição ou curso, ou então, que realizaram intercâmbio e não possuem registros de CR para os períodos cursados em outra instituição.

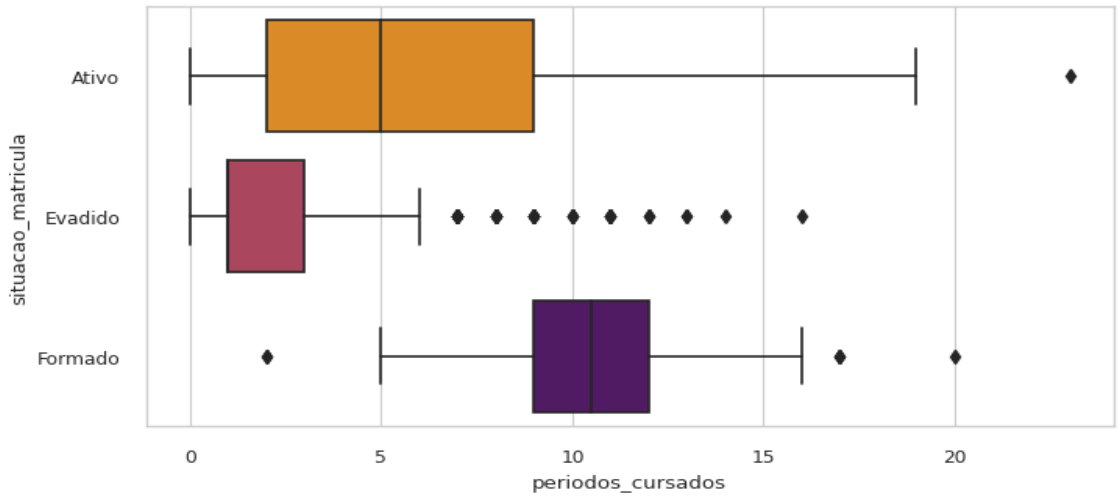
Além disso, percebe-se que os alunos que evadem de seus cursos o fazem nos primeiros períodos da graduação, visto que a distribuição do número de períodos cursados para estes alunos está majoritariamente concentrada entre 0 e 6 períodos. Sendo que aproximadamente 85% dos alunos evadem até o 5° período, 75% até o 3° período e 40% no primeiro período.

Para a construção dos modelos de aprendizado de máquina não serão considerados os alunos que possuem matrícula ativa. Essa decisão se deve a duas razões: (i) considerar alunos ativos como exemplos negativos de evasão aumenta muito a probabilidade de ruído nos registros, uma vez que alunos ativos possuem situação indefinida podendo se formar ou evadir; e (ii), conforme o trabalho de Solis et al. (2018), essa abordagem já demonstrou piores resultados.

4.3 Etapa 3: Preparação dos dados

A preparação dos dados proposta neste trabalho consiste na transformação dos dados, tratamento de dados faltantes e normalização das escalas dos atributos.

Figura 3: Número de períodos cursados por situação.



4.3.1 Transformação dos dados

Os algoritmos de aprendizado de máquina requerem que os dados estejam em um formato específico. Para isso, na maioria das aplicações, uma série de operações deve ser aplicada para transformar os dados do formato em que são encontrados naturalmente para outro que possa ser usado na construção dos modelos.

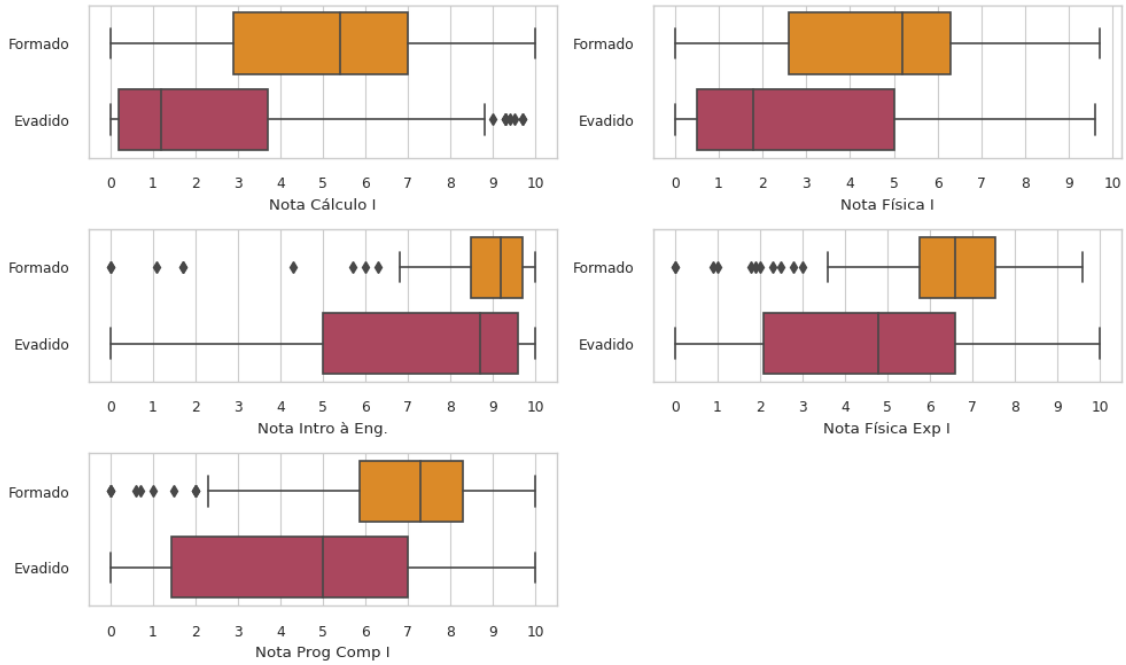
Os algoritmos considerados para este trabalho são desenvolvidos para trabalhar com dados no formato tabular. Neste formato, cada linha deve representar uma única instância (exemplo ou amostra) e cada coluna deve representar um atributo, normalmente numérico ou categórico, associado a essa instância. Esse processo também é conhecido como vetorização, pois efetivamente transforma cada instância em um vetor em espaço n- dimensional, onde n é o número de atributos considerados para a construção do modelo. No caso de uma tarefa de aprendizado supervisionado, como é a deste trabalho, cada instância possui também uma variável alvo, que o modelo tenta prever.

Para gerar uma representação única de cada aluno, é necessário que cada atributo represente um aspecto diferente do desempenho acadêmico do aluno e que a variável alvo seja o atributo relativo à situação da matrícula do aluno. Os dados disponíveis de cada aluno foram listados na Tabela 1.

Uma possível estratégia seria utilizar a nota dos alunos em todas as disciplinas cursadas. Entretanto, essa abordagem não foi utilizada devido a dois problemas. O primeiro é o conjunto de dados esparsos (com vários atributos sem valores), pois alunos dos diferentes cursos de Engenharia cursam disciplinas diferentes e algumas disciplinas não são obrigatórias para todos os alunos. Além disso, alunos dos primeiros períodos teriam valores nulos para todas as disciplinas dos próximos períodos. Para tratar esse problema, seria necessário usar alguma técnica para preencher os valores nulos, o que geraria grande ruído para alunos dos primeiros períodos. O segundo problema é o vazamento de informações, pois 40% dos alunos que evadem não chegam a completar o primeiro período, e acrescentar atributos dos demais períodos com notas vazias poderia evidenciar como um aluno que vai evadir e enviesar o modelo.

Assim, as informações usadas serão somente referentes ao desempenho do aluno no primeiro período: o coeficiente de rendimento, a nota e o resultado em cada disciplina. Enquanto o coeficiente de rendimento pode ser incluído diretamente como um atributo, os demais devem ser de alguma forma resumidos para serem considerados como atributos, para evitar a criação de um conjunto de dados esparsos. Para condensar o desempenho do aluno em uma métrica (diferente do coeficiente de rendimento), foi criado o atributo aproveitamento, que representa a proporção de disciplinas aprovadas no primeiro período, ou seja, a relação entre o número de aprovações e o número de disciplinas inscritas.

Figura 4: Comparação entre as distribuições de notas no 1º período de alunos formados e evadidos..



A Figura 4 mostra que, entre as disciplinas do primeiro período, aquelas que apresentam a maior diferença entre as distribuições de notas de alunos formados e evadidos são Cálculo I e Física I, especialmente em relação aos valores médios. Por isso, somente as notas dessas disciplinas no primeiro período serão consideradas como atributos para a construção do modelo.

Assim, o conjunto de dados final usado para a construção do modelo será composto por 10 atributos. Quatro deles são criados a partir dos resultados dos alunos nas disciplinas do primeiro período, enquanto os demais se referem ao período pré-admissão do aluno e são extraídos diretamente da base de dados. A Tabela 2 lista os atributos finais após essa etapa.

A decisão de utilizar esses atributos foi baseada na ideia de que o desempenho inicial em disciplinas fundamentais, como Cálculo I e Física I, pode ser um indicador da adaptação do aluno ao curso de engenharia. Como essas disciplinas envolvem habilidades matemáticas e raciocínio lógico, dificuldades nelas podem sinalizar desafios acadêmicos que influenciam a permanência no curso. Além disso, o modelo também inclui atributos relacionados ao desempenho prévio dos alunos no ENEM, conforme listado na Tabela 2. A junção desses dados permite uma análise mais abrangente, considerando tanto o histórico acadêmico antes do ingresso na universidade quanto o desempenho inicial no curso. Essa abordagem busca identificar padrões que possam auxiliar na

Tabela 2: Atributos utilizados na construção dos modelos.

Atributo	Descrição
nota_humanas	Nota da prova de Ciências Humanas do ENEM
nota_linguagem	Nota da prova de Linguagem do ENEM
nota_matematica	Nota da prova de Matemática do ENEM
nota_natureza	Nota da prova de Ciências da Natureza do ENEM
nota_redacao	Nota da prova de Redação do ENEM
cr	Coeficiente de rendimento do aluno em no primeiro período
aproveitamento	Taxa de aprovação nas disciplinas do primeiro
nota_fis_1p	Grau do aluno no primeiro período na disciplina de Física I
nota_calc_1p	Grau do aluno no primeiro período na disciplina de Cálculo I

predição da evasão e na implementação de estratégias preventivas.

Por fim, foi observado durante essa etapa que 101 alunos aparecem sem registros, sendo assim a base de dados ficou com 659 alunos, uma vez que os 692 alunos ativos também foram removidos, como informado na seção anterior.

4.3.2 Tratamento de valores faltantes

Alguns atributos possuem valores faltantes. A ausência dessas informações impacta no desempenho do algoritmo e precisa ser avaliada. A Tabela 3 apresenta o número de dados faltantes para cada atributo.

Tabela 3: Número de dados faltantes por atributos.

Atributo	Dados faltantes
nota_humanas	52
nota_linguagem	52
nota_matematica	16
nota_natureza	52
nota_redacao	16
cr	0
aproveitamento	0
nota_fis_1p	150
nota_calc_1p	102

As estratégias de tratamento de valores nulos utilizadas foram: (i) imputação da mediana; e (ii) imputação pelo algoritmo de agrupamento KNN. A mediana é uma técnica robusta contra outliers, garantindo que os valores imputados não sejam influenciados por extremos. Já o KNN, que leva em conta a similaridade entre observações, considera a suposição de proximidade entre os pontos e pode resultar em uma melhor estimativa dos valores ausentes.

4.3.3 Normalização das escalas dos atributos

Neste estudo, foi usado o algoritmo StandardScaler da biblioteca scikit-learn, que converte os valores de um conjunto de dados para uma escala comum, método chamado de scaling. Os valores resultantes são chamados de "dados padronizados" ou "normalizados", com média 0 e desvio padrão 1. O scaling costuma melhorar o tempo de convergência do modelo à solução e sua capacidade preditiva (Raju et al., 2020). Essa abordagem foi aplicada a todos os atributos do conjunto.

4.4 Modelagem

Na etapa de modelagem, o principal objetivo é encontrar a combinação de modelo e conjunto de hiperparâmetros que demonstre o melhor desempenho para o problema estudado.

4.4.1 Seleção dos modelos

Foram escolhidos 3 algoritmos de classificação que diferem em complexidade e abordagem, de modo a obter diferentes perspectivas do problema. Os modelos escolhidos foram:

- Regressão Logística
- Árvore de decisão
- eXtreme Gradient Boosting (XGBoost)

A regressão logística é um modelo linear, portanto é menos flexível e pode não ser capaz de capturar relações complexas não lineares nos dados (Izbicki & dos Santos, 2020). O modelo de árvore de decisão, por outro lado, é não-linear e pode capturar uma gama mais ampla de padrões nos dados. Entretanto, essa flexibilidade também pode tornar a árvore de decisão mais suscetível ao overfitting, especialmente se a árvore for muito profunda (com muitos níveis de nós de decisão).

Em geral, a regressão logística é um bom ponto de partida para tarefas de classificação, pois é um algoritmo simples e rápido de treinar. Por outro lado, embora as árvores de decisão possam ser mais poderosas, sua eficácia pode depender fortemente da escolha de hiperparâmetros, e elas podem exigir mais esforço para otimizar e interpretar (Hastie et al., 2009).

Já o XGBoost, por ser uma implementação de um algoritmo ensemble, que combina as previsões de múltiplas árvores de decisão simples e criadas sequencialmente tentando melhorar a previsão da anterior, consegue ser mais preciso que árvores de decisão únicas. Entretanto, devido ao seu maior poder preditivo, caso não tenha sua complexidade limitada, é um modelo bastante propenso ao overfitting (Mienye & Sun, 2022).

4.4.2 Otimização dos hiperparâmetros

Para a otimização dos hiperparâmetros e seleção dos melhores modelos de cada algoritmo, foi utilizada a biblioteca Optuna (Diaz-Pace et al., 2022), que recentemente se tornou a implementação de técnicas de otimização preferida da comunidade open source, devido à sua versatilidade, velocidade e eficiência em encontrar soluções ótimas quando comparada a outros algoritmos otimizadores (Banachewicz & Massaron, 2022; Shekhar et al., 2021).

Utilizando a biblioteca Optuna, foi escolhida uma estratégia de amostragem bayesiana implementada pela abordagem TPE - Tree-structured Parzen Estimator (Banachewicz & Massaron, 2022; Bergstra & Bengio, 2012), visto que o espaço de busca deste problema é bastante complexo, especialmente devido ao algoritmo XGBoost, que possui múltiplas vezes a quantidade de hiperparâmetros dos outros dois modelos.

Para a construção dos modelos, foram exploradas diferentes combinações entre algoritmos, estratégias de imputação de valores faltantes e ajustes de hiperparâmetros. Três algoritmos de classificação foram considerados: Regressão Logística, Árvore de Decisão e XGBoost. Para o tratamento de valores ausentes nos dados, duas estratégias foram testadas: imputação pela mediana e imputação baseada no algoritmo KNN.

Cada algoritmo teve seus principais hiperparâmetros ajustados com o objetivo de encontrar a melhor configuração de desempenho. No caso da Regressão Logística, foram ajustados os parâmetros penalty, que define o tipo de regularização aplicada, e C, que controla a intensidade dessa regularização. Para o algoritmo de Árvore de Decisão, os hiperparâmetros considerados incluíram criterion (critério utilizado para divisão dos nós), max_depth (profundidade máxima da árvore), min_samples_split (número mínimo de amostras exigido para dividir um nó) e min_samples_leaf (quantidade mínima de amostras em cada nó folha).

O algoritmo XGBoost, por sua vez, teve uma gama mais ampla de hiperparâmetros analisados, incluindo: eval_metric (métrica de avaliação), n_estimators (número de árvores a serem construídas), learning_rate (taxa de aprendizado), max_depth (profundidade máxima das árvores), max_delta_step (valor máximo de atualização permitido por nó), alpha e lambda (termos de regularização L1 e L2, respectivamente), gamma (ganho mínimo necessário para realizar uma divisão), min_child_weight (peso mínimo exigido para divisão de um nó), colsample_bytree (proporção de atributos utilizada por árvore), subsample (proporção de amostras utilizadas por árvore) e grow_policy (estratégia de crescimento das árvores).

A combinação dessas abordagens permitiu a avaliação de um espaço de busca robusto, balanceando simplicidade, capacidade de generalização e complexidade computacional. Esse detalhamento metodológico visa favorecer a reprodutibilidade do experimento e facilitar sua aplicação em estudos semelhantes.

Para encontrar a solução deste problema de otimização, a métrica utilizada para avaliação do modelo foi o valor de área debaixo da curva ROC, AUC. Sendo a função objetivo o valor médio desta métrica resultante da validação cruzada com 5 folds para cada combinação de hiperparâmetros. Além disso, para este processo foi utilizada uma amostra aleatória contendo apenas 80% dos dados do conjunto inicial, sendo os 20% restantes reservados para a inferência do desempenho final dos modelos.

O caso estudado neste trabalho é um problema de classificação desbalanceado, no qual a quantidade de exemplos de cada classe da variável objetivo não é a mesma (ver Figura 1). Isso pode ser um problema para a maioria dos algoritmos de aprendizado de máquina, que podem ser tendenciosos a prever desproporcionalmente instâncias da classe majoritária (Kuhn & Johnson, 2013). Para lidar com este problema, uma abordagem comum é utilizar estratégias de balanceamento por amostragem, como inclusão de amostras repetidas ou sintéticas (oversampling) ou exclusão do excesso de instâncias na classe majoritária (undersampling). Entretanto, ambas as estratégias já apresentaram problemas, principalmente em conjuntos de dados relativamente pe-

quenos ou com ruídos, ou quando são empregadas estratégias de imputação de valores nulos (Le et al., 2019; Zhihao et al., 2019).

Dessa forma, optou-se por uma estratégia mais simples, mas que produz resultados similares, sem os potenciais pontos negativos das técnicas de balanceamento por amostragem: atribuição de pesos inversamente proporcionais à quantidade de instâncias de cada classe. Desta forma, é atribuída uma importância inferior às informações obtidas das instâncias da classe majoritária, alterando assim a função de perda do algoritmo. Para isso, em cada algoritmo foi incluído um hiperparâmetro fixo de atribuição de pesos (class_weight = 'balanced'), o que é feito usando as opções do scikit-learn.

4.5 Etapa 5: Avaliação

Para avaliar o desempenho dos modelos de regressão logística, árvore de decisão e XGBoost após a etapa de otimização, foram usadas as métricas AUC, F1-Score, precisão e revocação, assim como métodos de avaliação gráfica.

Usando o procedimento de escolha de limiar de decisão, foram selecionadas as probabilidades de evasão acima das quais cada modelo classifica o aluno como "evadido". Assim, a capacidade de generalização dos modelos foi avaliada a partir da comparação de suas previsões com o resultado real dos alunos no conjunto de dados de teste.

É importante destacar que o presente estudo apresentou diversas limitações, como a baixa variedade dos atributos utilizados, a alta colinearidade entre os atributos e a qualidade dos dados utilizados. No entanto, acredita-se que os resultados deste trabalho são relevantes e podem ser generalizados para as turmas seguintes que ingressarão nos cursos de Engenharia no Instituto Politécnico da Universidade Federal do Rio de Janeiro.

5 Resultados

Na etapa de otimização de hiperparâmetros, foram encontradas as combinações de hiperparâmetros e processos de tratamento dos dados que retornaram o melhor valor de área debaixo da curva ROC (AUC). Essa métrica foi escolhida por ser estatisticamente robusta, principalmente frente a problemas de classificação desbalanceados. Nos três casos, a estratégia de imputação de valores nulos que retornou os melhores resultados foi a utilização do algoritmo KNN, sendo que, para os algoritmos da árvore de decisão e XGBoost foram considerados os valores de três vizinhos mais próximos e para a regressão logística de dez vizinhos mais próximos.

A Tabela 4 contém a avaliação do desempenho do melhor modelo de cada algoritmo. Estes resultados apontam que os três modelos escolhidos apresentaram desempenhos similares, com apenas alguns centésimos de diferença entre si e com intervalos de confiança equivalentes. Nota-se, no entanto, que o modelo da árvore de decisão apresenta médias de desempenho inferiores aos dois outros modelos.

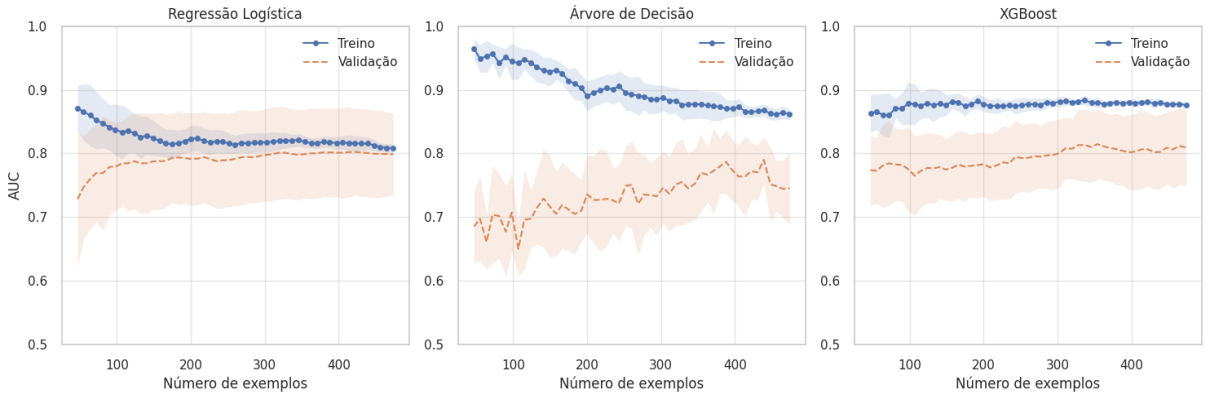
A Figura 5 mostra as curvas de aprendizado dos algoritmos utilizando a função learning_curve da biblioteca scikit-learn. Observa-se que o modelo da árvore de decisão apresenta considerável variância, evidenciada pela distância entre as curvas de treino e validação, além de um movimento

Tabela 4: Resultados dos modelos otimizados.

	AUC	**F1-Score**	**Precisão**	**Revocação**
Reg. Logística	0.800 ± 0.065	0.753 ± 0.053	0.733 ± 0.057	0.733 ± 0.057
Arv. de Decisão	0.753 ± 0.053	0.685 ± 0.069	0.681 ± 0.052	0.679 ± 0.054
XGBoost	0.811 ± 0.061	0.722 ± 0.066	0.721 ± 0.059	0.721 ± 0.059

ruidoso, indicando dificuldade do modelo de aprender os padrões contidos nos dados apresentados. Entretanto, percebe-se que, caso houvesse uma maior quantidade de dados, o desempenho do modelo continuaria a melhorar.

Figura 5: Curvas de aprendizado dos melhores modelos.



O modelo XGBoost também apresenta certa variância, evidenciada pela distância entre as curvas de treino e validação. Esse gap sugere que o modelo ainda não atingiu seu desempenho ideal na validação, o que pode indicar uma convergência mais lenta em comparação com a árvore de decisão. Esse comportamento é esperado, dado que o XGBoost é um modelo mais complexo, baseado em um conjunto de árvores construídas de forma sequencial, cada uma corrigindo os erros da anterior. Essa complexidade, aliada à regularização incorporada no modelo, torna seu aprendizado mais controlado e, por vezes, mais lento.

Contudo, observa-se que o XGBoost apresenta um comportamento mais estável, com menor oscilação entre diferentes execuções, o que é uma vantagem relevante para problemas com variabilidade nos dados. A tendência da curva de aprendizado do modelo indica que, com o aumento da quantidade de dados de treino, a diferença entre as curvas tende a diminuir. Isso ocorre porque mais dados permitem uma generalização melhor do modelo, reduzindo o sobreajuste e permitindo que os padrões relevantes sejam aprendidos com mais robustez. Além disso, a ampliação da base pode tornar mais efetivas as técnicas de regularização do XGBoost, contribuindo para uma convergência mais eficiente e precisa.

Já no caso da regressão logística, observa-se um comportamento praticamente ideal de uma curva de aprendizado: as curvas de treino e validação convergem simultaneamente a um mesmo valor da métrica de desempenho, com baixa variância.

Dessa forma, considerando apenas a curva de aprendizado, conclui-se que, dentre os modelos utilizados, a regressão logística seria o mais apropriado para o problema em questão, com a única ressalva de que, caso o algoritmo do XGBoost tivesse acesso a mais dados, é provável

que seu desempenho superasse o da regressão logística. Isto é de se esperar, dado o maior poder preditivo do XGBoost, capaz de extrair padrões mais complexos (não lineares, por exemplo), dos dados.

5.1 Escolha do Limiar de Decisão

A maioria dos modelos de classificação retorna como resultado de suas previsões um conjunto contendo a probabilidade de cada instância pertencer a cada uma das classes possíveis. Neste caso, como se trata de um problema de classificação binária, a probabilidade de uma instância pertencer a uma determinada classe é complementar à da classe oposta; assim, instâncias com probabilidade de pertencer à classe positiva igual ou superior a 0.5 são classificadas como positivas, caso contrário, como negativas. Entretanto, a distribuição desigual de probabilidade para dados desbalanceados faz com que os modelos tendam a prever mais instâncias da classe majoritária, porque têm mais dados dessa classe. Dessa forma, segue que a escolha de 0.5 como limiar de decisão não é suficiente para a garantia do melhor desempenho do modelo, sendo necessária a seleção de um limiar de decisão que permita o melhor desempenho de cada modelo.

Para a escolha dos limiares de decisão, foi utilizada a biblioteca ghostml, que implementa o método GHOST (Generalized tHreshOld ShifTing Procedure), apresentado por Esposito et al. (2021). Como métrica de avaliação de desempenho, foi utilizada a distância entre cada ponto da curva ROC e o canto superior do gráfico, ponto (0,1), que representa o desempenho de um modelo ideal. Dessa forma, para cada modelo e cada subconjunto de dados, é escolhido o limiar de decisão relativo ao ponto da curva ROC que mais se aproxima do canto direito do gráfico. Como resultado do procedimento, obtiveram-se os limiares de decisão que serão utilizados nas previsões finais dos modelos.

Tabela 5: Limiares de decisão obtidos pelo procedimento GHOST.

Modelo	**Limiar de decisão**
Regr. Logística	0.51
Árvore de Decisão	0.36
XGBoost	0.42

5.2 Previsões no conjunto de teste

Os limiares de decisão presentes na Tabela 5 foram utilizados para classificar as previsões dos modelos em relação aos dados do conjunto de teste, composto por 132 instâncias não utilizadas em nenhum momento anterior. As previsões dos modelos em relação ao conjunto de teste foram organizadas em matrizes de confusão, como mostra a Figura 6.

Ao comparar as matrizes de confusão dos diferentes modelos, foi observado que todos apresentaram desempenho semelhante, o que era esperado, considerando a proximidade de seus resultados no conjunto de dados de treinamento. O modelo de regressão logística obteve um número ligeiramente superior de verdadeiros positivos, classificando corretamente 53 alunos evadidos, um a mais que os demais modelos. A comparação dos desempenhos dos modelos utilizando outras métricas de avaliação está disponível na Tabela 6. Estes resultados confirmam o desempenho observado durante o treinamento.

Figura 6: Matrizes de confusão relativas às previsões dos modelos para as amostras do conjunto de teste.

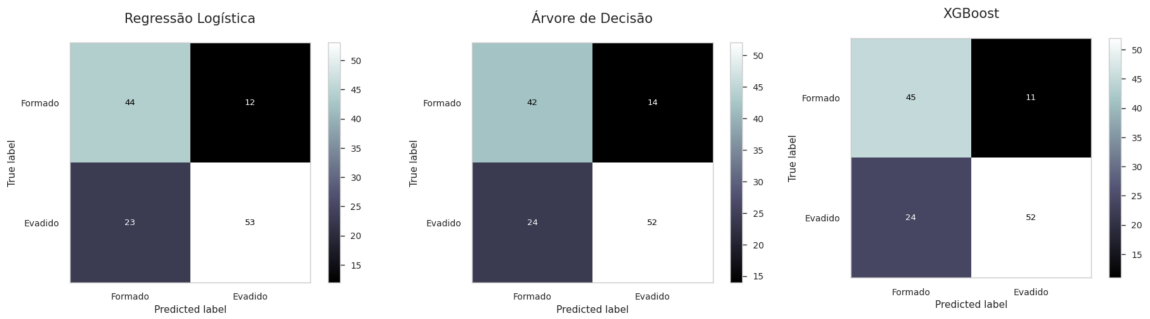
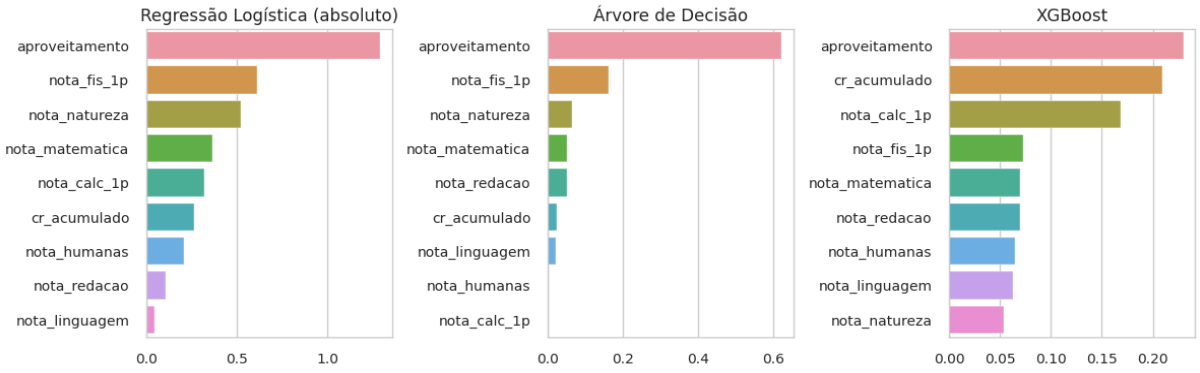


Tabela 6: Resultados dos modelos para o conjunto de dados de teste.

Modelo	**AUC**	**F1-Score**	**Precisão**	**Revocação**
Regressão Logística	0.805	0.752	0.735	0.697
Árvore de Decisão	0.759	0.732	0.712	0.684
XGBoost	0.803	0.748	0.735	0.684

A fim de melhor entender o funcionamento dos modelos em função do impacto de cada atributo nas suas previsões, foram utilizadas técnicas para cálculo de importância de atributos, específicas aos modelos. O resultado pode ser visto na Figura 7.

Figura 7: Importâncias de atributos intrínsecas aos modelos.



Como pode-se observar na Figura 7, o atributo aproveitamento é o mais importante para os 3 modelos, indicando que os alunos com baixo aproveitamento no primeiro período têm altas chances de evadir do curso. Para os modelos de regressão logística e árvore de decisão, os atributos mais importantes depois do aproveitamento são nota em física 1 e nota em ciências da natureza. É interessante observar que a nota em ciências da natureza engloba a disciplina de física do ensino médio, e essas duas notas podem indicar o aluno com dificuldades nessa área específica, o que impacta no seu desempenho nos cursos de Engenharia.

6 Discussão

Este estudo representa uma contribuição às iniciativas da Universidade Federal do Rio de Janeiro no combate à evasão universitária. A utilização de técnicas de ciência de dados e aprendizado de máquina para prever o risco de evasão e compreender melhor o desempenho dos estudantes fornece uma visão mais profunda sobre o fenômeno.

A análise dos dados revelou um alto índice de evasão nos primeiros 11 anos dos cursos de Engenharia, sendo um número 73% maior que a quantidade de alunos formados e que grande parte dos alunos das Engenharias evadem nos primeiros períodos, sendo 75% até o 3° período e 40% ainda no primeiro período, o que representa uma possibilidade de reaproveitamento dessas vagas. Além disso, quando comparados aos resultados da análise de sobrevivência de alunos feita por Klitzke e Carvalhaes (2023), nota-se que a taxa de evasão dos alunos em Macaé é superior àquela da universidade na totalidade, o que pode ser explicado pelo fato de ser um campus no interior do estado. Essa explicação é confirmada no trabalho de da Silva (2023) que indica relatos sobre a dificuldade em custear o transporte intermunicipal diariamente.

Os resultados da modelagem indicam que os atributos relacionados ao desempenho do aluno no primeiro período impactam sua probabilidade de evasão de forma mais significativa do que os atributos pré-matrícula. Em especial, o atributo “aproveitamento”, que mede a taxa de aprovação nas disciplinas do primeiro período, foi o atributo que mais contribuiu com as previsões de todos os modelos desenvolvidos. Além disso, as notas de física 1 e de ciências da natureza também foram importantes para dois dos três modelos construídos, indicando que essa é uma área que impacta na evasão dos alunos, quando eles não possuem uma boa base, o que também foi identificado nos achados de da Silva (2023), onde é dito que os estudantes chegam à universidade com "lacunas conceituais decorrentes de uma formação educacional básica deficiente".

As análises realizadas neste trabalho podem servir de guia para um melhor aproveitamento das vagas ao observar quando o aluno evade, assim como servir de referência para possíveis futuras políticas de combate à evasão implementadas pela universidade. No entanto, é importante destacar que as correlações identificadas neste estudo não são causais, e é necessária mais pesquisa para compreender completamente as interações complexas que influenciam a evasão escolar.

Este estudo considerou apenas atributos de desempenho acadêmico. Existem outros tipos de atributos que também podem influenciar o risco de evasão, como dados socioeconômicos, culturais e comportamentais. Incluir esses aspectos em análises descritivas e preditivas aumentaria a confiabilidade e a aplicabilidade dos resultados. A base de dados administrativos e acadêmicos da UFRJ é rica e pode ser usada e integrada na tomada de decisão.

Em relação ao desempenho dos modelos, o modelo baseado em regressão logística foi o mais apropriado para o problema em questão, com AUC 0,805, com a única ressalva de que, caso o XGBoost tivesse acesso a mais dados, é provável que seu desempenho superasse o da regressão logística.

Concluindo, este estudo fornece uma base sólida para trabalhos futuros no campo do Learning Analytics. O potencial de usar o aprendizado de máquina para prever o risco de evasão e compreender melhor o desempenho dos alunos é enorme para aperfeiçoar os resultados educacionais e aumentar a realização dos alunos.

Estudos futuros podem se concentrar no desenvolvimento de modelos mais sofisticados e na inclusão de fontes adicionais de dados, como informações demográficas dos estudantes ou feedback dos professores, para aumentar a acurácia e a interpretabilidade das previsões. Além disso, é importante considerar o impacto de intervenções específicas na redução do risco de evasão escolar. A implementação do modelo e a criação de um painel de monitoramento do desempenho dos alunos e do modelo são outras medidas recomendadas.

É crucial lembrar que, ao trabalhar com análise de dados e aprendizado de máquina, a governança e a privacidade dos dados são fundamentais para garantir a segurança dos dados.

Referências

Alharbi, M. (2020). The Economic Effect of Coronavirus (COVID-19) on Higher Education in Jordan: An Analytical Survey [Disponível em: Link]. *International Journal of Economics and Business Administration*, *8*, 521–532.

Aulck, L., Nambi, D., Velagapudi, N., Blumenstock, J., & West, J. (2019). *Mining University Registrar Records to Predict First-Year Undergraduate Attrition* (GS Search). International Educational Data Mining Society.

Banachewicz, K., & Massaron, L. (2022). *The Kaggle book: Data analysis and machine learning for competitive data science* (First edition) [GS Search]. Packt Publishing.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization [GS Search]. *Journal of Machine Learning Research*, *13*(10), 281–305.

Bielschowsky, C. E., & Amaral, N. C. (2022). O custo do aluno das 2.537 instituições de educação superior brasileiras: cai um mito? [GS Search]. *Educação & Sociedade*, *43*. https://doi.org/10.1590/ES.243866

Carvalhaes, F., Senkevics, A., & Costa Ribeiro, C. (2022). A interseção entre renda, raça e desempenho acadêmico no acesso ao ensino superior brasileiro [GS Search]. *Social Science Research Network*.

Chapman, P. (2000). CRISP-DM 1.0: Step-by-step data mining guide [Disponível em: Link].

Christo, M. M. S., Resende, L. M. M. d., & Kuhn, T. d. C. G. (2018). Por que os alunos de engenharia desistem de seus cursos – um estudo de caso [GS Search]. *Nuances: Estudos sobre Educação*, *29*. https://doi.org/10.32930/nuances.v29i1.4391

Coimbra, C. L., Silva, L. B. e., & Costa, N. C. D. (2021). A evasão na educação superior: definições e trajetórias [GS Search]. *Educação e Pesquisa*, *47*, e228764. https://doi.org/10.1590/S1678-4634202147228764

Crawford, J., Butler-Henderson, K., Rudolph, J., Malkawi, B., Glowatz, M., Burton, R., Magni, P. A., & Lam, S. (2020). COVID-19: 20 countries' higher education intra-period digital pedagogy responses [GS Search]. *Journal of Applied Learning and Teaching*, *3*(1), 09–28. https://doi.org/10.37074/jalt.2020.3.1.7

da Silva, I. C. (2023). *Mapeamento das experiências dos discentes dos cursos de Engenharia para redução da evasão: Um estudo de caso em uma Universidade Federal no interior do Rio de Janeiro* [Trabalho de Conclusão de Curso da Universidade Federal do Rio de Janeiro].

Diaz-Pace, J. A., Cian Berrios, R., Tommasel, A., & Vazquez, H. C. (2022). A Metrics-based Approach for Assessing Architecture-Implementation Mappings [GS Search]. *Anais Do*

XXV Congresso Ibero-Americano Em Engenharia de Software (CIbSE 2022), 16–30. https://doi.org/10.5753/cibse.2022.20960

Ebner, M., Schön, S., Braun, C., Ebner, M., Grigoriadis, Y., Haas, M., Leitner, P., & Taraghi, B. (2020). COVID-19 Epidemic as E-Learning Boost? Chronological Development and Effects at an Austrian University against the Background of the Concept of "E-Learning Readiness" [GS Search]. *Future Internet*, *12*(6), 94. https://doi.org/10.3390/fi12060094

Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. (2021). GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning [GS Search]. *Journal of Chemical Information and Modeling*, *61*(6), 2623–2640. https://doi.org/10.1021/acs.jcim.1c00160

Godoy, E. V., & Almeida, E. d. (2017). A evasão nos cursos de Engenharia e a sua relação com a Matemática: uma análise a partir do COBENGE [GS Search]. *Educação Matemática Debate*, *1*(3), 339–361. https://doi.org/10.24116/emd25266136v1n32017a05

Gomes, L. B. (2021). Elaboração de modelo de previsão da evasão universitária na Universidade Federal Fluminense através de métodos de aprendizado de máquina [Disponível em: Link].

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2ª ed.) [GS Search]. Springer.

Hernández-de-Menéndez, M., Morales-Menendez, R., Escobar, C. A., & Ramírez Mendoza, R. A. (2022). Learning analytics: state of the art [GS Search]. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *16*(3), 1209–1230. https://doi.org/10.1007/s12008-022-00930-0

Izbicki, R., & dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística* [GS Search].

Jesus, J. A. d., & Gusmão, R. P. d. (2024). Investigação da Evasão Estudantil por meio da Mineração de Dados e Aprendizagem de Máquina: Um Mapeamento Sistemático [GS Search]. *Revista Brasileira de Informática na Educação*, *32*, 807–841. https://doi.org/10.5753/rbie.2024.3466

Klitzke, M., & Carvalhaes, F. (2023). Fatores associados à evasão de curso na UFRJ: Uma análise de sobrevivência [GS Search]. *Educação em Revista*, *39*, e37576. https://doi.org/10.1590/0102-469837576

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* [GS Search]. Springer.

Le, T., Vo, M. T., Vo, B., Lee, M. Y., & Baik, S. W. (2019). A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction [GS Search]. *Complexity*, *2019*, 1–12. https://doi.org/10.1155/2019/8460934

Lobo, M. B. C. d. M. (2012). Panorama da evasão no ensino superior brasileiro: Aspectos gerais das causas e soluções [GS Search]. *Associação Brasileira de Mantenedoras do Ensino Superior*, 23.

Manhães, L. M. B., da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., & Zimbrão, G. (2012). Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados [GS Search]. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, *1*(1).

Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects [GS Search]. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Murphy, M. P. A. (2020). COVID-19 and emergency eLearning: Consequences of the securitization of higher education for post-pandemic pedagogy [GS Search]. *Contemporary Security Policy*, *41*(3), 492–505. https://doi.org/10.1080/13523260.2020.1761749

Nagy, M., & Molontay, R. (2018). Predicting Dropout in Higher Education Based on Secondary School Performance [GS Search]. *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, 389–394. https://doi.org/10.1109/INES.2018.8523888

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., & Moreira, F. (2021). How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review [GS Search]. *Big Data and Cognitive Computing*, *5*(4), 64. https://doi.org/10.3390/bdcc5040064

Oliveira, R. d. S., & Medeiros, F. P. A. d. (2024). Modelo de Predição de Evasão Escolar com Base em Dados de Autoavaliação de Cursos de Graduação [GS Search]. *Revista Brasileira de Informática na Educação*, *32*, 1–21. https://doi.org/10.5753/rbie.2024.3542

Oliveira Júnior, J. G. d. (2015, dezembro 8). *Identificação de padrões para a análise da evasão em cursos de graduação usando mineração de dados educacionais* [diss. de mestr., Universidade Tecnológica Federal do Paraná] [GS Search].

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification [GS Search]. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 729–735. https://doi.org/10.1109/ICSSIT48917.2020.9214160

Regehr, C., & Goel, V. (2020). Managing COVID-19 in a Large Urban Research-Intensive University [GS Search]. *Journal of Loss and Trauma*, *25*, 1–17. https://doi.org/10.1080/15325024.2020.1771846

Semesp. (2022). *Mapa do Ensino Superior* [Disponível em: Link]. Semesp.

Shekhar, S., Bansode, A., & Salim, A. (2021). A Comparative study of Hyper-Parameter Optimization Tools [GS Search]. *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 1–6. https://doi.org/10.1109/CSDE53843.2021.9718485

Silva, H. R. B., & Adeodato, P. J. L. (2012). A data mining approach for preventing undergraduate students retention [GS Search]. *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2012.6252437

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to Predict Dropout in University Students with Machine Learning [GS Search]. *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 1–6. https://doi.org/10.1109/IWOBI.2018.8464191

Teodoro, L. d. A., & Kappel, M. A. A. (2020). Aplicação de Técnicas de Aprendizado de Máquina Para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil [GS Search]. *Revista Brasileira de Informática na Educação*, *28*, 838–863. https://doi.org/10.5753/rbie.2020.28.0.838

Zhihao, P., Fenglong, Y., & Xucheng, L. (2019). Comparison of the Different Sampling Techniques for Imbalanced Classification Problems in Machine Learning [GS Search]. *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 431–434. https://doi.org/10.1109/ICMTMA.2019.00101